FACULTAD DE CIENCIAS AGRARIAS

UNIVERSIDAD NACIONAL DE ROSARIO

ALINEADO Y COMPARACIÓN DE SECUENCIAS GENÓMICAS OBTENIDAS DE GRUPOS DISCREPANTES PARA LA DETECCIÓN DE REGIONES CROMOSÓMICAS QUE CONTROLAN CARACTERES DE FRUTO EN TOMATE

Ing. Agr. Dana Valeria Vazquez

TRABAJO FINAL PARA OPTAR AL TÍTULO DE ESPECIALISTA EN BIOINFORMÁTICA

DIRECTOR: Ing. Agr. (Dr.) Vladimir Cambiaso

CO- DIRECTOR: Ing. Agr. (Dr.) Gustavo R. Rodríguez

AÑO 2019

ALINEADO Y COMPARACIÓN DE SECUENCIAS GENÓMICAS OBTENIDAS DE GRUPOS DISCREPANTES PARA LA DETECCIÓN DE REGIONES CROMOSÓMICAS QUE CONTROLAN CARACTERES DE FRUTO EN TOMATE

Dana Valeria Vazquez

Ingeniera Agrónoma – Universidad Nacional de Rosario

Este Trabajo Final es presentado como parte de los requisitos para optar al grado académico de Especialista en Bioinformática, de la Universidad Nacional de Rosario y no ha sido previamente presentado para la obtención de otro título en ésta u otra Universidad. El mismo contiene los resultados obtenidos en investigaciones llevadas a cabo en Zavalla, Argentina, durante el período comprendido entre marzo del 2017 y marzo del 2019 bajo la dirección del Dr. Vladimir Cambiaso y la Co-dirección del Dr. Gustavo R. Rodríguez.

_____

Nombre y firma del Director
Ing. Agr. (Dr.) Vladimir Cambiaso

_____

Nombre y firma del autor
Ing. Agr. Dana V. Vazquez

_____

Nombre y firma del Co – Director
Ing. Agr. (Dr.) Gustavo R. Rodríguez

Defendida: .....de 2019

Quiero agradecer en primer lugar a mis directores, el Dr. Vladimir Cambiaso y el Dr. Gustavo R. Rodríguez, por haberme propuesto realizar la Especialización en Bioinformática, guiarme y acompañarme a lo largo de todo el proceso. Gracias por su dedicación permanente, por la humildad y paciencia con la que me han enseñado y por apoyarme en cada etapa de este trabajo. Pero sobre todo por su gran calidad humana, y los aprendizajes que me han dado, mucho más allá de lo puramente académico. Porque con su ejemplo me incentivan a superarme como profesional y hacer lo que me apasiona. También, es mi deseo agradecer de forma especial al Dr. Javier Pereira da Costa, por la colaboración y la supervisión en mi trabajo, que ha realizado con gran generosidad y dedicación.

A todos los integrantes de la Cátedra de Genética de la Facultad de Ciencias Agrarias (UNR), por sus contribuciones contantes, por su predisposición y compartir humildemente su gran experiencia y conocimiento. Pero sobre todo por permitirme formar parte de este grupo, que es una familia. Quiero mencionar especialmente a la Dr. Roxana Zorzoli, porque su amor y dedicación en la docencia despertaron un sentimiento muy especial en mí y me contagiaron a iniciarme en la carrera Académica.

A mis compañeros de la "Sala de Becarios" por todos los momentos compartidos: las conversaciones, los almuerzos, las risas, y también las preocupaciones y frustraciones. Gracias a ustedes, porque hacen que todo sea más sencillo. A Meli, Agus, Pau, Luci, Baby, Mai, Shir, Maga, Mara y Cele, gracias por amistad, porque es lo más invaluable que me ha dado este trabajo.

A mis compañeros de la Especialización, por las largas horas de cursos, mates y almuerzos. Porque más de una vez he sentido que me "hablaban en chino" y gracias a su gran apoyo superé esos momentos, y he logrado aprender cosas que nunca habría imaginado.

Muchas gracias a mis docentes de la Especialización en Bioinformática, gracias a cada uno por su dedicación y paciencia. Porque es un gran desafío enseñar informática en dos años a alguien con formación en biología y sin embargo han hecho todo lo posible para darme las herramientas necesarias y han abierto todo un nuevo mundo para mí.

A mi familia y amigos, por su amor y apoyo en cada paso de mi vida. Gracias a mi hermana Dani, por su amor infinito, por ser mi pilar en la vida, por saber que decirme en cada oportunidad. Y especialmente a Bruno, mi gran compañero, gracias por el apoyo en cada una de mis decisiones, por la paciencia, gracias por ser mi profesor particular y por motivarme siempre, pero por sobre todo por el amor y el día a día. A todos ellos, gracias totales.

Finalmente gracias a la Fundación Ciencias Agrarias, por financiar este posgrado.

PUBLICACIONES Y PRESENTACIONES A CONGRESOS

IDENTIFICACIÓN IN SÍLICO DE NUEVAS REGIONES GENÓMICAS ASOCIADAS AL CARÁCTER TIPO DE CARPELO EN FRUTOS DE TOMATE (S. LYCOPERSICUM). Vazquez, D. V., Cambiaso, V., Pereira da Costa, J. H., Rodríguez, G. R. Trabajo presentado en el "XVII Congreso Latinoamericano de Genética, XLVII Congreso Argentino de Genética, LII Reunión Anual de la Sociedad de Genética de Chile, VI Congreso de la Sociedad Uruguaya de Genética, V Congreso Latinoamericano de Genética Humana y V Simposio Latinoamericano de Citogenética y Evolución (ALAG/2019)". 6 al 9 de octubre de 2019, Mendoza, Mendoza, Argentina

ABREVIATURAS Y SÍMBOLOS

ADN: ácido desoxirribonucleico.

***BSA*:** (*Bulked-segregant analysis*) análisis de grupos segregantes.

FDR: (*False Discovery Ratio*) tasa de falso descubrimiento.

ha: hectárea.

***InDel*:** Inserción-Deleción. Es una inserción o eliminación de una o más bases nucleotídicas.

MAS: (*Marker-assisted Selection*) selección asistida por marcadores

Mb: mega bases (un millón de bases).

***NGS*:** (*Next-Generation Sequencing*) secuenciación de última generación.

pb: pares de bases.

***QTL*:** (*Quantitative Trait Locus*) locus de carácter cuantitativo.

***RIL*:** (*Recombinant Inbred Lines*) líneas recombinantes endocriadas.

***SNP*:** (*Single Nucleotide Polymorphism*) polimorfismo de nucleótido único.

tn: tonelada.

RESUMEN

El tomate es un cultivo modelo ampliamente utilizado en los estudios de mejoramiento vegetal, que posee una gran cantidad de recursos e información genómica y posgenómica disponibles. En el año 2012 se obtuvo el primer genoma de referencia en tomate a partir del cultivar Heinz 1706 de *Solanum lycopersicum* L. y hasta la fecha se han secuenciado y alineado a esta referencia más de 700 genomas pertenecientes a especies silvestres y cultivadas. La morfología de los frutos de tomate es un aspecto de interés ya que define el destino de la producción y preferencia en el mercado de consumo en fresco, y por lo tanto el estudio de las bases genómicas que controlan este carácter y su aplicación en programas de mejoramiento tiene un impacto económico directo. La técnica *QTL-seq* es una metodología para la identificación de regiones genómicas asociadas a caracteres de interés agronómico en forma rápida y eficiente. Esta técnica combina el análisis de grupos segregantes (*BSA* o *Bulked-segregant analysis*) y la secuenciación de genomas completos. En este proyecto se propone alinear las secuencias genómicas de dos grupos de plantas de tomate que difieren para el carácter tipo de carpelo y detectar los polimorfismos asociados a dicho carácter mediante la implementación de la técnica *QTL-seq.* En la generación F_2 del cruzamiento entre los cultivares de tomate *(S. lycopersicum L.)* 'Voyage', que presenta frutos con carpelos no fusionados y 'Old Brooks' que tiene frutos con carpelos fusionados se seleccionaron 10 plantas con frutos fusionados y 10 con frutos no fusionados. Se realizó la extracción de ADN de las 20 plantas y se mezcló en partes iguales para obtener grupos de ADN segregantes para el carácter tipo de carpelo. Las muestras de ADN se secuenciaron y alinearon a las versiones SL2.50 y SL3.0 del genoma de referencia de tomate. Se compararon las secuencias alineadas entre sí obteniendo una lista de polimorfismos entre los grupos segregantes respecto a la secuencia de referencia. Se utilizó la metodología G´ del paquete *QTLseqr* de R para detectar las regiones genómicas subyacentes al rasgo de interés. Se detectaron tres regiones con comportamiento diferencial entre ellos, en los cromosomas 3, 6 y 10. Estas regiones controlarían el carácter tipo carpelo en la población segregante analizada. La región del cromosoma 3 presentó una longitud de 2,88 Mb y un valor máximo de G´ de 4,69 en la posición 56,48 Mb (FDR (q)<0,05), la región del cromosoma 6 abarcó 9 Mb, y tuvo un valor máximo de G´ de 9,8 en la posición 43,57 Mb (FDR (q)<0,01); mientras que la región del cromosoma 10 fue la más extensa, con 59 Mb, y un valor máximo de G´ de 7,27 en la posición 27,95 Mb. (FDR (q)<0,01). Estos resultados serán complementados con estudios genéticos para identificar los mecanismos subyacentes al tipo de carpelo en fruto y diseñar marcadores moleculares para implementar en programas de mejoramiento del cultivo de tomate.

PALABRAS CLAVE: ANÁLISIS *IN SILICO*, GENÉTICA VEGETAL, MORFOLOGÍA DE FRUTO*, QTL-seq*, *Solanum lycopersicum*, TIPO DE CARPELOS.

ABSTRACT

WHOLE-GENOME SEQUENCE ALIGNMENT AND COMPARISON FROM DISCREPANT GROUPS FOR DETECTION OF CHROMOSOMIC REGIONS THAT UNDERLIE TOMATO FRUIT TRAITS

Tomato is one of the most important worldwide cultivation. It is also a model crop widely used in plant breeding studies. There are also many resources available as well as genomic and post genomic information. The fruit morphology in tomato is an aspect of agronomic importance because it defines the production destination and the consumer preferences. Therefore, the study of the genomic bases that control this trait and its application in breeding programs has a direct economic impact. The QTL-seq technique combines the analysis of segregating groups (BSA or Bulked-Segregant Analysis) and the whole-genome sequencing in order to identify genomic regions associated with traits of interest. In this project, we propose to align the genomic sequences from two groups of tomato plants that differ for type of carpel and implement the *QTL-seq* technique to detect polymorphisms associated with that trait. We based on an F_2 population derived from the cross between the tomato cultivars (*S. lycopersicum* L.) 'Voyage' which presents fruits with unfused carpels, and 'Old Brooks' that has fused fruits. Ten plants with fused and unfused fruits were selected. DNA extraction from the twenty plants was performed and mixed to obtain DNA from two segregating groups for type of carpel. The DNA samples were sequenced and aligned to the SL2.50 and SL3.0 versions of the tomato genome reference. The aligned sequences were compared to each other to obtain a list of polymorphisms. The G´ methodology was implemented using the R package named *QTLseqr* to detect the genomic regions underlying the trait of interest. It was possible to determine three putative regions, at chromosomes 3, 6, and 10 that would control the type of carpels. The detected G´ values were 4.69 (FDR (q) <0.05), 9.8 and 7.27 (FDR (q) <0.01) respectively. The whole-genome sequences of two discrepant groups for type of carpels were aligned to the tomato reference and three genomic regions with a differential allele distribution between the groups, associated to the trait of interest using the QTL-seq technique were detected. These results will be complemented with genetic studies to identify the mechanisms underlying type of carpel in fruits and to design molecular markers that can be implemented in tomato breeding programs.

ÍNDICE

INTRODUCCIÓN

El cultivo de tomate es una de las hortalizas más consumidas en el mundo, y presenta gran relevancia en el sector agrícola, siendo el octavo cultivo más importante en producción (FAOSTAT, 2017). La producción mundial actual se estima en 182.301.395 tn, el área sembrada es de 4.848.384 ha y el rendimiento promedio de 37,6 tn/ha (FAOSTAT, 2017). Por su parte en nuestro país el consumo de tomate es de 16 kg/habitantes/año, superando ampliamente la media mundial (5,6 kg/hab/año), la producción nacional es de 660.753 tn, y el área cosechada representa 16.915 ha, lo que equivale a un rendimiento de 39 tn/ha (FAOSTAT, 2017). En la provincia de Santa Fe se encuentra el Cinturón Hortícola de Rosario, ubicado en la zona de influencia de la Facultad de Ciencias Agrarias de la Universidad Nacional de Rosario. Este se destaca por el nivel de producción, comercialización, área sembrada e ingresos brutos que genera (Censo INTA, 2012).

El tomate cultivado es diploide en su constitución genética, con un número básico de 12 cromosomas, un genoma de tamaño pequeño (900 Mb) y un ciclo de cultivo corto. Estas características biológicas junto a la disponibilidad de recursos genéticos y bases de datos con información genómica y posgenómica lo convierten en uno de los cultivos modelos más efectivos para el mejoramiento.

A partir del cultivar Heinz 1706 de *Solanum lycopersicum* L. se obtuvo la secuencia completa del genoma de referencia en tomate (The Tomato Genome Consortium, 2012). Dos grandes proyectos de re-secuenciación se han llevado a cabo luego de la obtención del genoma de referencia (Aflitos et al., 2014; Lin et al., 2014), alienándose a la misma referencia hasta el momento más de 700 accesiones de tomate. Recientemente, se ha construido el primer pan-genoma en tomate a partir de la información genómica de 725 accesiones silvestres y cultivadas seleccionadas por ser filogenética y geográficamente representativas de la diversidad existente (Gao et al., 2019). Toda la información de las distintas -ómicas del tomate y de otras especies pertenecientes a la familia de las Solanáceas se encuentra almacenada y disponible públicamente en el sitio web de la red genómica de Solanáceas (The Sol Genomics Network, http://solgenomics.net; Fernandez-pozo et al., 2015; Mueller et al., 2005).

La morfología del fruto en tomate es una característica de gran importancia económica, histórica y biológica. Desde el punto de vista económico, la forma del fruto define el destino de la producción (mercado en fresco o industria) y la elección del consumidor en el mercado en fresco. Desde el punto de vista histórico, ha sido un criterio de selección durante la domesticación y el mejoramiento del cultivo. Y desde el punto de vista biológico, la morfogénesis es un mecanismo complejo y diverso entre los distintos tipos de órganos y entre las especies. El fruto en las especies silvestres de tomate es esférico y de pequeño tamaño. Durante el proceso de domesticación y mejoramiento el fruto de tomate

evolucionó a formas y tamaños muy variables. La gran diversidad para la forma del fruto de tomate es mayormente explicada por mutaciones presentes en sólo cinco genes: *SUN, OVATE, SOV1, LOCULE NUMBER* (*LC*) *y FASCIATED* (*FAS*). Los alelos mutados o cultivados en *SUN, OVATE y SOV1* (Liu et al., 2002; Rodríguez et al., 2011a Rodríguez et al., 2013; Wu et al., 2018; Xiao et al., 2008) controlan la forma alargada de los fruto; mientras que los alelos cultivados de *FAS* (Cong et al., 2008) y *LC* (Muños et al., 2011) dan lugar a frutos más aplanados y mayor número lóculos (cavidades internas de los frutos en las que se ubican las semillas). Dependiendo del contexto genético en el que se encuentren, el efecto de los alelos cultivados en los genes *FAS* y *LC* es variable sugiriendo que estos interactúan con otros genes para expresar el fenotipo final. A su vez, el alelo mutante de *FAS* se asocia no solo con más carpelos sino también con carpelos no fusionados (Tanksley, 2009). El tamaño de frutos está definido principalmente por el número de carpelos presentes en la flor y lóculos que forman el fruto. Se sabe que un aumento en el número de carpelos y lóculos puede llevar a un incremento de hasta 50% en el tamaño final del fruto (Fernández-Lozano et al., 2014; Lippman y Tanksley, 2001; Tanksley, 2004). Entonces la morfología, tamaño y cantidad de lóculos son caracteres de fruto relacionados entre sí, y el estudio de las bases genéticas subyacentes a los mismos resulta de gran interés.

Los polimorfismos constituyen variaciones a nivel de secuencia genómica en una región específica del ADN y generan las distintas alternativas alélicas para un gen en una población. Aquellos polimorfismos que afectan a la secuencia codificante o reguladora y que producen cambios importantes en la estructura de la proteína o en el mecanismo de regulación de la expresión, pueden dar lugar a cambios en el fenotipo o la manifestación de un carácter. Por su parte estos polimorfismos pueden constituir una sustitución de una única base nitrogenada, los cuales se denominan polimorfismos de nucleótido único (*SNP*, del inglés *Single Nucleotide Polymorphism*) o bien una inserción o eliminación de una o más bases, denominado Inserción-Deleción (*InDel*). Los marcadores moleculares de ADN permiten identificar polimorfismos a nivel de secuencia entre individuos o genotipos. Cuando los marcadores moleculares se encuentran próximos a un gen de interés, existe nula o muy baja recombinación entre ellos, es decir que están estrechamente ligados y por lo tanto siempre se heredan juntos. En consecuencia es posible utilizarlos como un indicador de la presencia de una región de interés (Collard et al., 2005).

Los caracteres de interés agronómico son en su mayoría caracteres complejos, controlados por muchos genes, cada uno con un efecto aditivo relativamente pequeño sobre el carácter (Falconer y Mackay, 1996). Este tipo de caracteres se denominan cuantitativos mientras que a aquella región genómica asociada, o con efecto significativo, sobre un carácter cuantitativo se la denomina *locus* de carácter cuantitativo (*QTL,* del inglés *Quantitative Trait Locus*). Tradicionalmente los *QTL* han sido detectados mediante análisis de ligamiento en una población derivada de un cruzamiento entre progenitores con fenotipos contrastantes para un carácter de interés. Sin embargo, en estos cruzamientos muchos *QTL*

podrían estar segregando dificultando identificar el efecto de un *locus* particular. Con las nuevas tecnologías de secuenciación y la posibilidad de utilizar *SNP* como marcadores moleculares se han podido utilizar como progenitores individuos más relacionados genéticamente. En los cruzamientos derivados de estos, el efecto fenotípico de los genes menores se puede expresar como genes mayores en la población, siendo más fácilmente identificable cuando uno o pocos *loci* que controlan el carácter está segregando. Por lo tanto, en poblaciones segregantes derivadas del cruzamiento entre cultivares divergentes para los caracteres de interés y en las que los genes mayores no presentan polimorfismo es posible identificar nuevas regiones genómicas que controlan el carácter o bien regiones que previamente han sido informadas con un efecto menor sobre el mismo (Rodríguez et al., 2013). De este modo, para descubrir la totalidad de genes que controlan la diversidad de la forma en el tomate cultivado, análisis genéticos y moleculares deben ser conducidos en poblaciones segregantes obtenidas a partir de cruzamientos intraespecíficos o entre cultivares.

El análisis de grupos segregantes (*BSA*, del inglés *Bulked Segregant Analysis*) constituye una enfoque alternativo para la detección de *QTL* (Giovannoni et al., 1991; Mansur et al., 1993; Michelmore et al., 1991). Este consiste en construir una población segregante (F_2 o líneas recombinantes endocriadas (*RIL*, del inglés *Recombinant inbred lines*)) y fenotiparla, para seleccionar individuos con valores extremos para el rasgo de interés. El ADN de estos individuos se mezcla para constituir grupos (*bulks*) de fenotipos contrastantes y se efectúa su genotipado. El principio básico en que se sustenta es que en la mayoría de las regiones genómicas, la proporción de los genomas parentales deben ser aproximadamente iguales en ambos *bulks*, excepto en las regiones estrechamente asociadas con el carácter de interés, en donde estarán sobre o sub representado uno u otro genoma parental de manera diferencial entre los *bulks*.

Esta metodología se aplicó por primera vez para evaluar el comportamiento en poblaciones F_2 de caracteres discretos como son la abscisión pedicular (*jointless*) y maduración del fruto (*non-ripening*) en tomate (Giovannoni et al., 1991) y la resistencia a enfermedades en lechuga (Michelmore et al., 1991). En un principio el genotipado se realizó mediante el uso de marcadores moleculares (Asnaghi et al., 2004; Giovannoni et al., 1991; Michelmore et al., 1991) y posteriormente este concepto se extendió al análisis de caracteres cuantitativos (Chagué et al., 1997; Quarrie et al., 1999; Zhang et al., 2009). Finalmente con el desarrollo de la tecnología de secuenciación de última generación (*NGS*, del inglés *Next-Generation Sequencing*) de ADN se proporcionaron herramientas efectivas para la detección de *SNP* en todo el genoma (Huang et al., 2009), evitando la necesidad de desarrollar marcadores moleculares para mapear. Por lo tanto, realizar el *BSA* mediante secuenciación de grupos discrepantes, constituye una herramienta valiosa ya que reduce el tiempo y labor requeridos en la identificación de *QTL* al realizar el genotipado solo de aquellos individuos con fenotipos extremos en vez de una población completa de mapeo, y

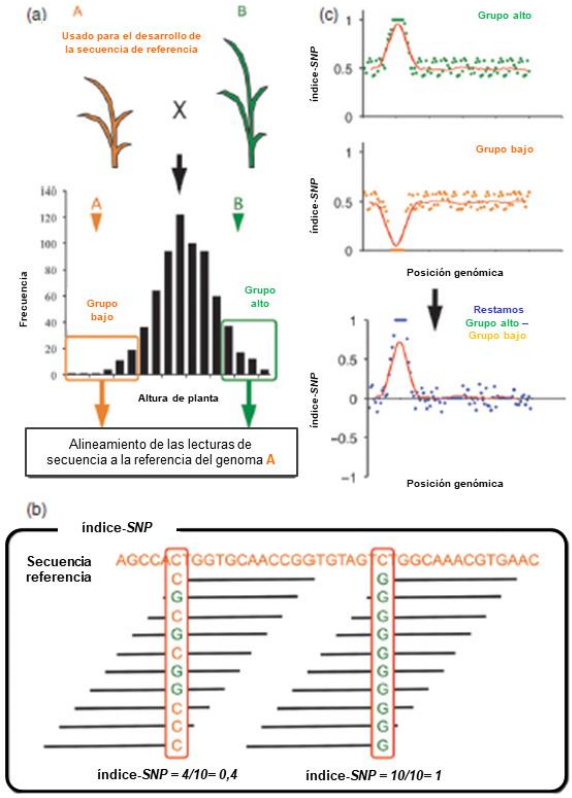
evitar la necesidad de desarrollar marcadores moleculares específicos para esta tarea. En la actualidad esta técnica se ha aplicado incluso en estudios de RNA-seq (Wang et al., 2013) y proteínas (Pereira da Costa et al., 2017).

La disponibilidad de los recursos genómicos y bioinformáticos han permitido implementar el enfoque de *BSA* basado en *NGS* con éxito para identificar *QTL* asociados a peso de fruto y número de lóculos en tomate (Illa-Berengue et al., 2015) y mapear distintos *QTL* en numerosas especies, como por ejemplo: maíz (Chen et al., 2017), arroz (Takagi, Abe, Yoshida, Kosugi, Natsume, Mitsuoka, & Uemura, 2013), garbanzo (Das et al., 2015), pepino (Lu et al., 2014) y soja (Song et al., 2017).

La principal metodología utilizada para realizar *BSA* basado en la secuenciación de genomas completos, llamada *QTL-seq*, fue propuesta por Takagi et al. (2013) y ha sido ampliamente empleada. Este enfoque permite reconocer regiones genómicas en la que los progenitores difieren y que contribuyen a los mayores o menores valores del carácter de interés en la progenie resultante, posibilitando así detectar *QTL* asociados al carácter (Takagi, Abe, Yoshida, Kosugi, Natsume, Mitsuoka, & Uemura, 2013). Primero se genera una población de mapeo a partir de cruzamiento entre cultivares con fenotipos contrastantes para el carácter de interés y se realiza el fenotipado. Cuando existen varios *loci* implicados en la variación fenotípica, la distribución de frecuencias tiende a una normal. Se seleccionan aquellos individuos con fenotipos extremos para conformar dos grupos o *bulks* discrepantes para el carácter (Figura 1a). Se extrae el ADN de cada planta de cada grupo y se mezclan proporcionalmente para conformar una muestra por grupo. La muestra proveniente de cada grupo debe ser secuenciada y alineada al genoma de referencia de la especie. Se espera que cada *bulk* presente el genoma de ambos padres en una relación 1:1 para la mayoría de las regiones del genoma y una desigual representación de los genomas de los progenitores en las regiones asociados a las diferencias fenotípicas entre los grupos, es decir los *QTL*. Para determinar la cantidad relativa de genoma derivado de cada padre, se evalúa la cantidad de lecturas cortas correspondiente a cada uno de los genomas parentales para cada una de las posiciones donde se detecta un *SNP*. Para ello luego de alinear la secuencia de cada *bulk* con la secuencia de referencia de cualquiera de los parentales, se contabiliza la cantidad de lecturas que contienen un *SNP* y se define la proporción de lecturas alternativas en relación a la cobertura total, lo que se denomina índice-*SNP* (Figura 1b). Cuando el índice-*SNP* es igual a 0, indica que todas las lecturas son como la secuencia del padre de referencia, mientras que un índice-*SNP* de 1 indica una sobre representación del genoma del otro padre. Por su parte un índice-*SNP* de 0,5 indicaría una igual contribución de los genomas parentales en la progenie. Luego los datos de ambos *bulks* se combinan restando el valor de índice-*SNP* del *bulk* menor al valor del mayor, obteniendo un nuevo parámetro denominado Δ(índice-*SNP*). Se espera que los *bulks* sean idénticos para aquellas regiones que no contribuyen a las diferencias fenotípicas entre ambos, con un valor de índice-*SNP* igual a 0,5 y Δ(índice-*SNP*) de cero. Mientras que en las regiones que contienen los *QTL*, ambos *bulks*

tienen una representación desigual del genoma parental por lo tanto los valores índice-*SNP* serán opuestos y el Δ(índice-*SNP*) tendrá un valor absoluto de uno, en función del parental que aporta los alelos que contribuyen a las diferencias fenotípicas el valor será positivo o negativo. Es decir, si el *bulk* de mayor valor presenta los alelos alternativos, el valor de índice-*SNP* será igual a uno y el otro *bulk*, que presenta los alelos como la referencia, tendrá un valor de índice-*SNP* igual a cero, por lo tanto el valor de Δ(índice-*SNP*) será igual a 1; en el caso contrario el valor será -1. Finalmente se promedian y grafican los Δ(índice-*SNP)* y se calcula mediante un proceso de simulación estadística un umbral de intervalo de confianza. Las regiones con un valor absoluto de Δ(índice-*SNP*) que sobrepasan este umbral poseen polimorfismos significativamente diferenciales entre los *bulk,* y por lo tanto son regiones candidatas para contener *QTL*.

Figura 1: Esquema adaptado de la metodología de *QTL-seq* desarrollada en arroz por Takagi et al. (2013).



a) Los fenotipos fueron cruzados para generar una progenie F_2 que se segrega para altura de planta. En este ejemplo, el padre A tiene baja estatura mientras que el padre B tiene alta estatura. Dado que muchos *loci* de rasgos cuantitativos (*QTL*) controlan este carácter, la frecuencia en la progenie F_2 sigue una distribución normal. Se seleccionaron las plantas con la estatura más alta y más baja, y se mezcló el ADN para formar el grupo "alto" y "bajo", respectivamente. El ADN de estos grupos se secuenció y alineó con la secuencia de referencia del cultivar A para calcular el índice-*SNP*. (b) Definición del índice-*SNP*. Las lecturas cortas generadas por la secuenciación del genoma completo se alinearon con la secuencia de referencia. Si 10 lecturas cortas cubren una posición de nucleótido dada, la cobertura del sitio es 10 X. Entre las 10 lecturas, si cuatro contienen un *SNP* diferente del nucleótido de referencia, el índice-*SNP* se define como 0,4. Por otro lado, si todas las

lecturas albergan un *SNP* diferente de la referencia, el índice-*SNP* es 1,0. (c) Ejemplos de gráficos de índice-*SNP*. El *QTL* puede identificarse como picos o valles del gráfico de índice-*SNP*. Cada punto corresponde a un *SNP*, y el eje x corresponde a la posición cromosómica. Las líneas sólidas se corresponden con valores promedio de índice-*SNP* o Δ(índice-*SNP*) en función de la posición genómica. Arriba: gráfico de índice-*SNP* del grupo "alto". Medio: gráfico del índice-*SNP* del grupo "bajo". Abajo: una gráfica de Δ(índice-*SNP*).

Por su parte Magwene et al. (2011) plantea un enfoque alternativo para evaluar estadísticas significativas y detectar *QTL* utilizando *BSA* a partir de la secuenciación de lecturas cortas de alto rendimiento. Esta metodología se basa en un valor suavizado del estadístico G para cada *SNP*, ponderado de acuerdo a la distancia a un *SNP* focal. El método propuesto consiste primero en calcular el valor del estadístico G para cada *SNP*, que considera las frecuencias alélicas observadas y las esperadas suponiendo que la profundidad de lectura es igual para cada alelo en ambos *bulks*. Esta estadística presenta una distribución χ^2, sin embargo debido al esquema de muestreo jerárquico esto no se cumple, aun bajo la hipótesis nula (no existen *QTL* cercano al *SNP* considerado o *SNP* focal). Una importante fuente de variación es el efecto de margen aleatorio. Para lidiar con dicha variación se utiliza un promedio ponderado de G en función de sus *SNP* vecinos, llamado G´. Esto se fundamenta en que la divergencia en la frecuencia de alelos entre *bulks* se conserva entre sitios estrechamente ligados, pero la variación aleatoria debida a la cobertura de lecturas de secuenciación variable no. Así G´ constituye la sumatoria de todos los valores de G para los *SNP* contenidos dentro de una ventana deslizante de tamaño específico, ponderados por su distancia al *SNP* focal. Este tipo de media móvil ponderada, donde los pesos están dados por una función núcleo, también se conoce como regresión del núcleo de Nadaraya-Watson (Nadaraya, 1964; Watson, 1964). Esta regresión actúa como una función de suavizado, y la cantidad de suavizado aumenta con el tamaño de ventana deslizante (Schucany, 2004). Para ponderar los distintos *SNP* se utiliza la función núcleo de tricubo, que considera la distancia de cada *SNP* al *SNP* focal, dentro de la ventana y otorga mayor importancia a las observaciones que están cerca del *SNP* focal, es decir a medida que la distancia al *SNP* focal aumenta el factor de ponderación se acerca a 0. Luego, realiza la estimación no paramétrica de la distribución nula de G´, para esto se asume que en los datos observados hay regiones con distribución nula (no *QTL*) y otras una distribución contaminante (*QTL*). Se sabe que la distribución nula es sesgada a la derecha con una densidad de cola razonablemente explicada por una función log normal. Para calcular los valores medios y desvíos, primero es necesario eliminar las regiones contaminantes o valores atípicos (*QTL*). Así mediante la regla de Hampel (Hampel, 1978) se identifican los probables valores atípicos de un conjunto de datos y se extraen las distribuciones contaminantes, formando un conjunto de datos recortados. Finalmente, la media y la varianza del conjunto se estiman utilizando la mediana y el modo y los valores de significancia se calculan a partir de una distribución normal de este nuevo conjunto de datos. Por ultimo mediante un enfoque de tasa de descubrimiento falso (FDR; *False Discovery Ratio*) se estima valor umbral y se identifican los sitios que se desvían significativamente de la distribución nula de fondo. El estadístico suavizado de G

permite reducir la varianza a la vez que considera el desequilibrio de ligamiento en los *SNP*. Además, dado que el estadístico G´ presenta una distribución cercana a una log normal, es posible estimar un valor de significancia o p-valor para cada *SNP* usando una estimación no paramétrica de la distribución nula de G´(Mansfeld y Grumet, 2018).

Mansfeld y Grumet (2018) desarrollaron un paquete de herramientas para el programa R (RCore Team, 2016) llamado *QTL-seqr* que implementa ambas metodologías, Δ(índice-*SNP*) y G´, para el análisis de datos provenientes del *BSA* por secuenciación y la rápida detección de regiones genómicas asociadas a rasgos de interés.

Los cultivares de tomate (*S. lycopersicum* L.) 'Voyage' y 'Old Brooks' se caracterizan por poseer frutos de gran tamaño y alto número de lóculos y se diferencian en caracteres de morfología tales como la irregularidad de la superficie externa y en la fusión de los carpelos que formaron el fruto. Es remarcable que 'Voyage' presenta un fenotipo único con carpelos no fusionados. Se sabe que ambos cultivares presentan los alelos mutados en los genes *LC* y *FAS* y los alelos silvestres o *wild type* en los genes *SUN*, *OVATE* y *SOV1*, sin embargo esto no es suficiente para explicar sus características morfológicas y la discrepancia para el carácter tipo de carpelo. Por su parte en las poblaciones derivadas del cruzamiento entre estos dos cultivares, los genes mayores no presentarán polimorfismos, pero sí segregarán otros genes que controlan caracteres de morfología. Por consiguiente, en dichas poblaciones sería posible identificar nuevas regiones genómicas que controlan caracteres de morfología. Se ha observado que el carácter tipo de carpelo segregó en forma discreta en las distintas poblaciones derivada del cruzamiento entre 'Voyage' y 'Old Brooks', ajustándose a lo esperado para un carácter monogénico (resultados no publicados), es decir que un gen estaría determinando el carácter. Por su parte el fenotipo no fusionado se comporta como recesivo. En función de este resultado, nos centramos en el carácter tipo de carpelo para llevar a cabo la secuenciación de grupos segregantes.

A partir de estos antecedentes, en este proyecto nos proponemos analizar y comprar las secuencias genómicas completas de grupos discrepantes para el carácter tipo de carpelos a partir de la F_2 del cruzamiento entre 'Voyage' y 'Old Brooks' e implementar la técnica *QTL-seq* a fin de detectar nuevas regiones genómicas que controlan dicho carácter.

OBJETIVO GENERAL

Alinear las secuencias genómicas de dos grupos de plantas de tomate que difieren para el tipo de carpelo y detectar los polimorfismos asociados a este carácter.

OBJETIVOS ESPECÍFICOS

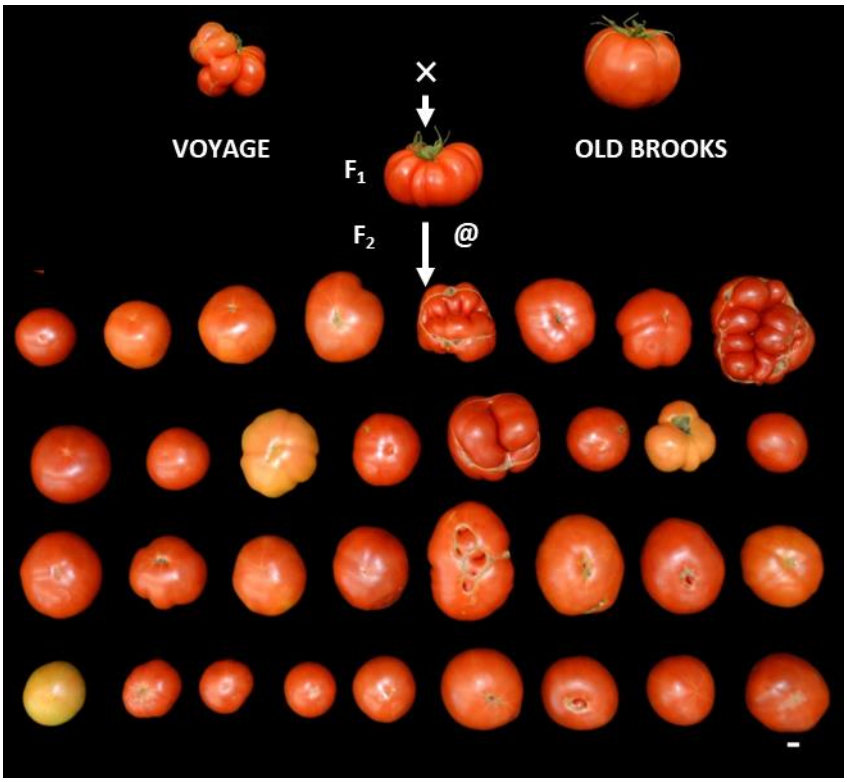
1) Alinear respecto del genoma de referencia en tomate la secuencia genómica de los grupos discrepantes para el carácter tipo de carpelo y obtener una lista de polimorfismos a partir de la comparación de las secuencias previamente alineadas.
2) Identificar aquellos polimorfismos genómicos asociados al carácter tipo de carpelo.

MATERIALES Y MÉTODOS

Material vegetal:

Se utilizaron los cultivares de tomate (*S. lycopersicum* L.) 'Voyage' y 'Old Brooks' como progenitores iniciales. Por cruzamiento manual se obtuvo el híbrido 'Voyage' X 'Old Brooks' y por autofecundación de la F_1 se generó la población F_2 segregante de 76 individuos (Figura 2). Se analizaron caracteres de morfología en las generaciones uniformes (parentales e híbrido) y la población F_2. Se seleccionaron 20 plantas F_2 con fenotipos discrepantes para el carácter tipo de carpelo, es decir 10 plantas con frutos fusionados y 10 plantas con frutos no fusionados, para conformar dos grupos o *bulks* contrastantes para el carácter de interés.

Figura 2: Esquema del proceso de obtención del material vegetal e imágenes de frutos representativos de los progenitores y distintas generaciones. Escala: 1cm.



Extracción y secuenciado de muestras de ADN:

Para reconstruir el genotipo de estas plantas F_2 seleccionadas se sembraron 24 semillas de cada planta F_2 para construir familias F_2:F_3. Las semillas se germinaron y sembraron en multimacetas con sustrato N°1 de Dynamics (www.dynamicscompost.com.ar), se mantuvieron en cámara con condiciones de temperatura y fotoperíodo controladas hasta que alcanzaron aproximadamente un tamaño de cuatro hojas verdaderas. Se extrajo una

hoja de cada planta de la familia F_2:F_3 y se mezclaron para conformar una única muestra representativa del genoma de la planta F_2 de origen. El ADN genómico fue aislado de hojas jóvenes siguiendo las metodologías descriptas por Bernatzky y Tanksley (1986) y Fulton et al.(1995). Se mezcló proporcionalmente el ADN de cada planta formando dos grupos (*bulks*). El ADN de cada grupo se llevó a una concentración 50 ng/ul. Las muestras de los dos grupos se enviaron a secuenciar en el Centro de Acceso a la Tecnología del Genoma o GTAC (Universidad de Washington, Saint Louis, Missouri, EEUU) con un equipo Illumina HiSeq 2500 en una célula de flujo de 2x101 lecturas apareadas.

Alineado de secuencias genómicas:

Se usó la herramienta Trim Galore! versión 0.4.5 (Babraham Bioinformatics Group, 2017) para eliminar las lecturas de baja calidad y cortar los adaptadores de los archivos *FASTQ*. Primero se cortaron los adaptadores Illumina (secuencia universal AGATCGGAAGAGC). Luego se filtraron las porciones de las lecturas con baja calidad, considerando como valor límite un Q=20, lo que equivale a un 99% de precisión. Para controlar la rigurosidad del proceso de eliminación del adaptador, se especificó una superposición requerida con la secuencia del adaptador mínima de 6 bases. Finalmente, se eliminaron ambas lecturas apareadas si al menos una de las dos fue más corta que 60 pb. Una vez obtenidos los archivos *FASTQ* validados y recortados, estos se alinearon con las versiones SL2.50 (liberación 2014) y la SL3.0 (liberación 2017) (https://solgenomics.net/organism/Solanum_lycopersicum/genome) del genoma de referencia en tomate utilizando la versión 2.3.2 de la herramienta Bowtie 2 (Langmead y Salzberg, 2012). Durante el alineado, se usó la opción "--very-sensitive-local" la cual define una serie de parámetros predeterminados de velocidad, sensibilidad y precisión del proceso. La opción "sensitive" implica que el proceso es más lento pero más sensible y preciso, mientras que la opción "-local" permite cortar algunos caracteres de uno o ambos extremos si esto logra maximizar la puntuación de alineación. Como resultado se obtuvieron archivos de secuencia alineados, en formato *SAM* (mapa de alineación de secuencia o en inglés *Sequence Alignment Map*). Este es un formato basado en texto que almacena información de alineación de secuencias cortas a una secuencia de referencia. Estos archivos se organizaron por coordenadas, se clasificaron, se etiquetaron y se convirtieron al formato *BAM* (mapa de alineación binaria o en inglés *Binary Alignment Map*) utilizando la versión 1.119 del programa Picard (http://picard.sourceforge.net). Este archivo es el equivalente binario de un archivo SAM que almacena los mismos datos pero en una representación binaria comprimida. Con este programa, se identificaron y etiquetaron las lecturas duplicadas. Además se creó un archivo de diccionario de secuencias (con la extensión ".dict") a partir de una secuencia de referencia en formato *FASTA*, ya que éste es requerido en otras etapas del proceso y un archivo índice para el *BAM* de entrada, lo que permite una búsqueda rápida de datos. Los archivos *BAM* de salida se analizaron con la versión 2.2.1 de

Qualimap (García-Alcalde et al., 2012). Este programa permite obtener estadísticas y gráficos que facilitan el control de calidad de los datos de secuencia de alineación y otras características como la cobertura obtenida en todo el genoma.

Comparación de secuencias alineadas:

Para comparar las secuencias genómicas alineadas de ambos *bulks* se obtuvieron para cada muestra, archivos que contienen las variantes genómicas llamados *gVCF* por las siglas en ingles de *Genomic Variant Calling Format*. Se utilizó la herramienta *HaplotypeCaller* de la versión 4.0.9.0 del programa GATK (DePristo et al., 2011; McKenna et al., 2010). Estos archivos presentan información de secuencia para todas las posiciones del genoma, indicando si las mismas son o no variantes del genoma de referencia. Luego se usó la herramienta *CombineGVCFs* de GATK para combinar los archivos *gVCFs* de ambos *bulks* en un único archivo *gVCF*. La función *GenotypeGVCFs* en GATK se usó para realizar el genotipado conjunto de ambas muestras incluidas en el archivo *gVCF*, generando un archivo en formato *VCF* (*Variant Call Format*) que contiene las variantes (*InDel* y *SNP*) detectadas luego de la comparación de las secuencias. Los *SNP* e *InDel* se extrajeron en archivos *VCF* separados mediante la opción "--select-type-to-include" de la herramienta *SelectVariants* de GATK. A continuación se filtraron los *SNP* de baja calidad o poco confiables con la opción *VariantFiltration* de GATK. Por último se eliminaron todos los sitios que no pasaron los filtros de calidad con la opción "--remove-filter-all" de la herramienta *VariantFiltration* incluida en la versión 0.1.15 del programa VCFtools (Danecek et al., 2011).

El proceso de alineado y comparación de las secuencias genómicas se realizó utilizando el Centro de Computo de Alto Rendimiento perteneciente al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) Rosario.

Identificación de regiones genómicas asociadas al carácter tipo de carpelo:

Mediante la herramienta *VariantsToTable* del programa GATK (Auwera et al., 2013) se extrajeron del archivo *VCF* combinado, para cada posición en cada uno de los *bulks*, la siguiente información: cromosoma, posición física, alelo de referencia y alternativo, cantidad de lecturas para cada alelo, la profundidad de cobertura total, genotipo detectado en cada *bulk* y valores de calidad. Esta herramienta devuelve una tabla de datos donde cada fila contiene la información descripta arriba para cada uno de los polimorfismos detectados (Tabla S1).

Se utilizó la versión 0.7.4 del paquete de R llamado *QTL*seqr (Mansfeld y Grumet, 2018) que incorpora, los métodos descriptos por Takagi et al. (2013) y Magwene et al. (2011) para realizar el análisis de grupos segregantes o *BSA*. Esta herramienta permite importar y filtrar los datos de *SNP* provenientes del análisis con herramientas de GATK para calcular y graficar la distribución de *SNP*, los valores de Δ(índice-*SNP),* G´ suavizados por tricubo y

valores de -$\log_{10}$ (p-valor), para la rápida identificación de los *QTL* asociados al carácter de interés. Se utilizó la función *importFromGATK* del programa *QTL*seqr para importar la tabla de datos obtenida anteriormente y calcular la frecuencia del alelo de referencia (1), el índice-*SNP* (2) *y el* Δ(índice-*SNP*) (3), para cada *SNP*.

$$Frecuencia\ del\ alelo\ de\ referencia = \frac{Profundidad\ del\ alelo\ ref\ _{Bulk\ no\ fusionado} + Profundidad\ del\ alelo\ ref\ _{Bulk\ fusionado}}{Cobertura\ total\ de\ ambos\ bulks} \quad (1)$$

$$índice - SNP_{por\ bulk} = \frac{Profundidad\ del\ alelo\ alternativo}{Profundidad\ total} \quad (2)$$

$$\Delta(\text{índice} - \text{SNP}) = índice\_SNP_{bulk\ no\ fusionado} - índice\_SNP_{bulk\ fusionado} \quad (3)$$

Con la función *filterSNP* se filtraron los *SNP* en función de la calidad y la profundidad de cobertura. Para ello se calcularon y consideraron los descriptores estadísticos (valores mínimos y máximos, media, mediana, primer y tercer cuartil) y se analizó también la información en forma gráfica. Se eliminaron aquellos *SNP* con una profundidad de cobertura en ambos *bulks* menor a 38 pb (primer cuartil) o mayor a 65 pb (cuarto cuartil), debido a que los polimorfismos detectados con baja cobertura son poco confiables y aquellos que presentan un excesivo valor de cobertura probablemente sean causados por dificultades en el alineado de regiones repetitivas del genoma. Se filtraron los *SNP* con una frecuencia del alelo de referencia menor a 0,2 o mayor a 0,8, quitando *SNP* bajo o sobre representados en ambos *bulks.* Se removieron los *SNP* con una discrepancia de profundidad de cobertura de hasta 50 pb entre los *bulks* y aquellos con una profundidad de lectura por *bulk* menor a 5 pb. Finalmente se aplicó un filtro por calidad, conservando solo aquellos *SNP* con una probabilidad de error menor al 1%.

Luego se llevó a cabo el análisis de los grupos segregantes propiamente dicho. Para esto primero se realizó el análisis de *QTL-seq* con la función *runQTLseqAnalysis* del paquete de R *QTL*seqr*.* El análisis se basa en calcular las diferencias de frecuencia de los alelos o Δ(índice-*SNP*), a partir de las profundidades de los alelos en cada *SNP.* Durante este análisis la función *runQTLseqAnalysis* realiza una serie de pasos: primero cuenta la cantidad de *SNP* presentes en una ventana de pb definida por el usuario, luego se calcula el valor de Δ(índice-*SNP*) y se suaviza por tricubo dentro de la ventana ($tricubo\_\Delta\ (\text{índice} - SNP)(4)$). Es decir cada valor de Δ(índice-*SNP*) se multiplica por un factor (kj) que considera la distancia relativa desde el *SNP* focal, que tiene un valor máximo en el centro de la ventana y un valor mínimo en el extremo (5). A continuación se calcula la mínima profundidad de lectura en cada posición y la profundidad suavizada por tricubo dentro de la ventana; y finalmente se calcula el intervalo de confianza mediante un método de simulación derivado de los datos de profundidad de lectura. Esta simulación consiste en determinar la probabilidad de muestrear

un alelo alternativo, de acuerdo a la profundidad de lectura y la estructura poblacional (F_2 o RIL) y se calcula el índice-*SNP* para cada *bulk*. Esto se repitió mediante proceso de muestreo por iteración 10.000 veces y se calculó el de Δ(índice-*SNP*) restando el *bulk* simulado B al *bulk* simulado A. Los valores extremos de los cuantiles (99 o 95%) de estos 10.000 Δ(índice-*SNP*) representan casos raros en la replicación y por lo tanto son estimaciones de los intervalos de confianza. Estos valores de intervalo simulados se relacionan al dato de profundidad mínima correspondiente a cada ventana deslizante y se calcula el intervalo de confianza en la ventana. Una región se considera significativamente diferente de 0, si el valor absoluto del $tricubo\_\Delta\,(\text{índice} - SNP)$ es mayor al intervalo de confianza en esa ventana deslizante.

$$tricubo\_\Delta\,(\text{índice} - SNP) = \sum_{j\ en\ W} k_j * \Delta\,(\text{índice} - SNP) \quad (4)$$

kj: núcleo tricubo: D_J es la distancia relativa de un *SNP j* a un *SNP* focal, con el valor 0 en la posición focal y el valor 1 en el borde de la ventana deslizante W.

$$kj = \frac{\left(1-D_j^3\right)^3}{S_W} \quad (5)$$

S_W: es la suma de $\left(1 - D_j^3\right)^3$ para todos los *SNP* en la ventana deslizante W.

Luego se realizó el enfoque alternativo desarrollado por Magwene et al. (2011) con la función *runGprimeanalysis* del paquete de R *QTL*seqr, para determinar la significancia estadística de los *QTL*. Este método calcula una estadística G modificada para cada *SNP* en función de las profundidades observadas de los alelos y las esperadas, que se calculan asumiendo que la profundidad de lectura es igual para todos los alelos en ambos grupos (6). Este valor de G se multiplica por un valor de ponderación núcleo tricubo, en una ventana cuyo tamaño en pb depende de las tasas de recombinación conocidas o estimadas (7). El valor de G' resultante constituye un estadístico G suavizado con tricubo mediante regresión local constante dentro de cada cromosoma.

$$G = 2\sum_{i=1}^{4} n_i \ln\left(\frac{Obs(\,n_i\,)}{Esp\,(n_i)}\right) \quad (6)$$

n_i : profundidad del alelo de referencia y alternativo para cada *bulk,* con un valor de i de 1 a 4*;* $Obs(\,n_i\,)$ es la profundidad de alelo observada y $Esp\,(n_i)$ son los valores esperados.

$$G' = \sum_{j\ in\ W} kj\ Gj \quad (7)$$

W : es la ventana deslizante, kj: núcleo tricubo, Gj: valor de G para el *SNP j*

A su vez esta función estima un valor de significancia o p-valor para cada *SNP* con una estimación no paramétrica de la distribución nula de G' descripta por Magwene et al. (2011). Se usa el enfoque de tasa de descubrimiento falso, FDR, que considera las comparaciones múltiples y se estima un umbral de p-valor apropiado (o el umbral de G correspondiente)

para determinar los sitios que se desvían significativamente de la distribución nula. Este valor de significancia se ajusta mediante Benjamini-Hochberg (Benjamini y Hochberg, 1995) obteniendo un Q-valor. Entonces se puede considerar que los *QTL* candidatos son aquellas regiones continuas y/o sitios que superan ese valor umbral. Se calculó también el $-\log_{10}$ (p-valor), ya que es una forma de visualizar más fácilmente los *QTL*.

Esquematización de *QTL* putativos y exportación de datos:

Se graficó el Δ(índice-*SNP*) junto a los intervalos de confianza (95 y 99% confianza) para todas las posiciones del genoma, mediante las opciones *var = "deltaSNP*" y *plotIntervals = TRUE* de la función *plotQTLStat* del paquete de R *QTL*seqr (Mansfeld y Grumet, 2018). Las regiones que presentan Δ(índice-*SNP*) superiores al umbral de confianza constituyen *QTL* putativos.

También se graficaron los valores de G' y FDR (q) o umbral de 0,01 y 0,05 de significancia a lo largo de todo el genoma. Esto se realizó utilizando las opciones *var = "Gprime"* y *plotThreshold = TRUE* de la función *plotQTLStat*. Aquellas regiones con valores de G' que superan el umbral de FDR constituyen las regiones genómicas subyacentes al carácter de interés.

Finalmente, con la opción *var = "negLog10Pval"* de la función *plotQTLStat* se graficó el *-log$_{10}$ (p-valor)* derivado de G' en relación a la posición genómica, solo para los cromosomas donde encontramos *QTL* putativos.

La información de los *QTL* putativos se exportó en formato CSV con la función *GetQTLTable.*

Comparación de datos fenotípicos de los grupos segregantes para el carácter cuantitativo Grado de Irregularidad Externa:

Puesto que la metodología fue desarrollada para caracteres cuantitativos y en el análisis previo se aplicó a un carácter del tipo cualitativo, determinado por el efecto de un *locus* o pocos *loci*, se decidió probar la metodología para un carácter cuantitativo discrepante entre los grupos ya conformados. Por lo tanto, se evaluó la existencia de diferencias significativas entre los valores medios de los *bulks* para el carácter grado de irregularidad externa de los frutos.

Para determinar el grado de irregularidad externa, primero se escanearon cortes transversales de ocho frutos promedio por cada planta. Se siguieron las recomendaciones propuestas por Rodríguez et al. (2011b). Las imágenes digitales se guardaron en formato JPG. Estas se analizaron con el programa Tomato Analyzer 3.0 (Rodríguez et al., 2010; Rodríguez et al., 2011b) que calcula automáticamente y en forma cuantitativa la irregularidad externa de los frutos. Para calcular el grado de irregularidad externa el programa realiza una serie de radios desde el centro del fruto al pericarpio y calcula el desvío estándar de las distancias. En los frutos circulares, los radios tienen aproximadamente

la misma longitud, y por lo tanto el desvío y grado de irregularidad tiende a cero, mientras que en los frutos muy irregulares el valor de desvío es mayor.

Se analizó la distribución normal de los datos de grado de irregularidad, mediante la prueba de Shapiro-Wilk (Shapiro y Wilk, 1965) y en forma gráfica por medio de un histograma. Se realizó un análisis *t de Student*, para grado de irregularidad utilizando como variable de clasificación el carácter tipo de carpelos. Se consideró un nivel de significancia del 5%.

RESULTADOS Y DISCUSIÓN

Extracción y secuenciado de muestras de ADN:

Se obtuvieron para cada grupo de ADN dos archivos con la secuencia genómica total fragmentada en millones de lecturas cortas, uno de cada extremo de la molécula de ADN. Estos datos fueron almacenados en archivos *FASTQ,* este formato se basa en texto y contiene los datos de secuencia de nucleótidos y sus correspondientes puntuaciones de calidad.

Alineado y comparación de secuencias genómicas:

Mediante el alineado de los datos de secuenciación a la versión SL2.50 (liberación 2014) y la SL3.0 (liberación 2017) del genoma de *Solanum lycopersicum* y la comparación genómica de los polimorfismos hallados en los dos *bulks* respecto a la referencia, fue posible identificar una serie de *SNP* e *InDel* distribuidos a lo largo de todo el genoma que difieren entre los grupos segregantes.

En la Tabla 1 se detalla la cantidad de polimorfismos identificados entre ambos grupos respecto a la secuencia de referencia, en los alineados correspondientes a las versiones SL2.50 y SL3.0 del tomate cultivado. En la alineación respecto a la versión SL2.50 del genoma de referencia fue posible identificar 206.431 *SNP* y 140.995 *InDel* a lo largo de todo el genoma. Al analizar su distribución en los distintos cromosomas se observó la mayor cantidad de *SNP* en el cromosoma 4 y de *InDel* en el cromosoma 1. Por su parte la menor cantidad de *SNP* se detectó en el cromosoma 7 mientras que la de *InDel* en el cromosoma 8. Luego de alinear ambos *bulks* respecto a la versión SL3.0 del genoma de referencia, se identificaron 207.674 *SNP* y 129.094 *InDel*. A diferencia del alineado a la versión SL2.50, la mayor cantidad de *SNP* y de *InDel* se detectaron en el cromosoma 10 y en el 1, respectivamente. Por su parte, los cromosomas 7 y 8 presentaron el menor número de *SNP* e *InDel*, al igual que en la versión SL2.50 del genoma de referencia. Los polimorfismos se encuentran distribuidos a lo largo de todo el genoma en los cromosomas, y el número de polimorfismos no parece estar relacionado con la longitud del cromosoma (Tabla 1).

La cantidad de polimorfismos encontrados entre ambos *bulks* respecto al genoma de referencia resulta mucho menor que los valores encontrados en un trabajo similar realizado por Cambiaso et al.(2019) (1.081.626 *SNP* y 315.892 *InDel*). Esto podría explicarse ya que en dicho trabajo se realiza la comparación del genoma de una especie cultivada y una especie silvestre alineadas respecto a la secuencia de referencia, mientras que en nuestro análisis efectuamos la comparación de los extremos fenotípicos en una población segregante derivada del cruzamiento de dos cultivares. Por lo tanto, es esperable que las diferencias genómicas sean mayores entre especies más distanciadas genéticamente. Sin embargo la

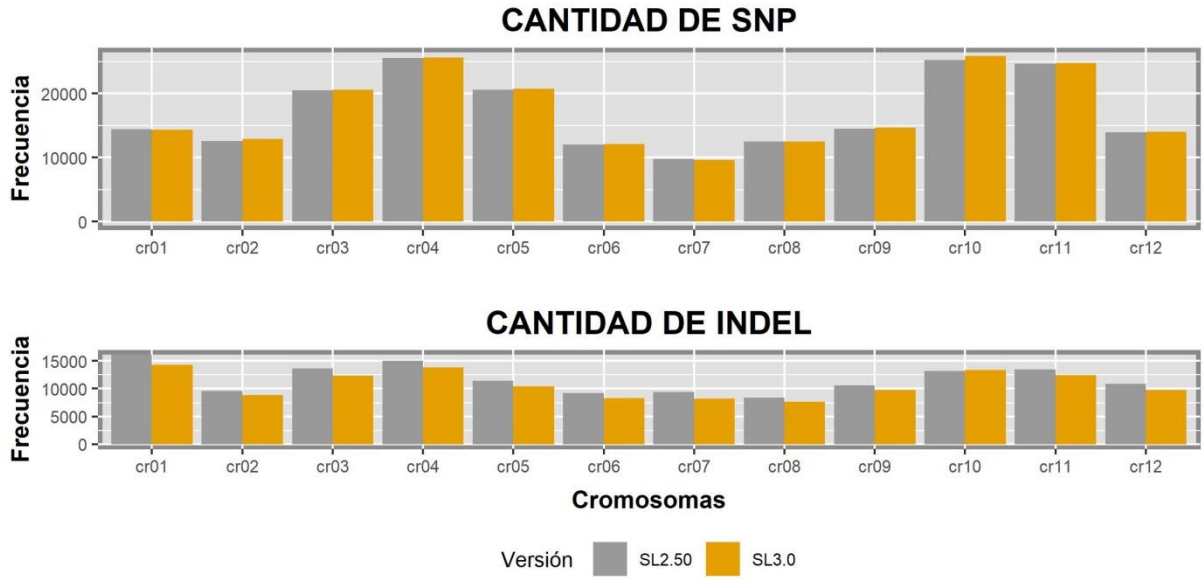
cantidad de polimorfismos detectados puede variar ampliamente de acuerdo a las metodologías utilizadas y materiales analizados (Causse et al., 2013; Jung et al., 2016; Yang et al., 2014).

Tabla 1: Cantidad de polimorfismos de tipo *SNP* e *InDel* identificados entre ambos bulks respecto de la secuencia de referencia (versiones SL2.50 y SL3.0) por cromosoma (Cr) y su respectiva longitud en mega bases (Mb).

	SL2.50			**SL3.0**		
	Longitud (Mb)	***SNP***	***InDel***	**Longitud (Mb)**	***SNP***	***InDel***
Cr01	98,54	14.425	16.139	98,46	14.342	14.263
Cr02	55,34	12.619	9.562	55,98	12.907	8.852
Cr03	70,79	20.526	13.636	72,29	20.603	12.357
Cr04	66,47	25.562	14.990	66,56	25.629	13.804
Cr05	65,88	20.557	11.454	66,72	20.716	10.429
Cr06	49,75	12.035	9.249	49,79	12.081	8.249
Cr07	68,05	9.779	9.391	68,18	9.637	8.199
Cr08	65,87	12.514	8.410	65,99	12.480	7.596
Cr09	72,48	14.519	10.626	72,83	14.631	9.788
Cr10	65,53	25.230	13.207	65,57	25.887	13.343
Cr11	56,30	24.692	13.463	56,60	24.727	12.445
Cr12	67,15	13.973	10.868	68,06	14.034	9.769
TOTAL	**823,94**	**206.431**	**140.995**	**827,87**	**207.674**	**129.094**

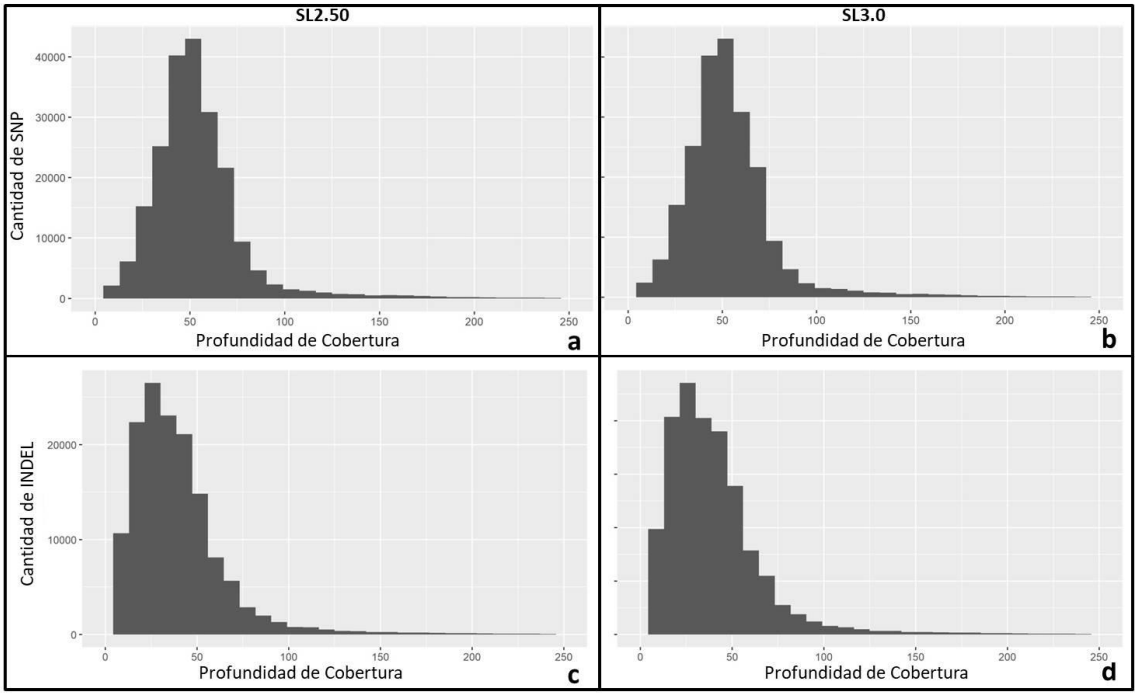
La distribución de la cantidad de polimorfismos detectados en cada cromosoma en ambas versiones del genoma de referencia se representa en la Figura 3. Se puede observar que la cantidad total de *SNP* identificados aumentó en el alineado respecto de la versión más reciente del genoma SL3.0. Sin embargo, estas diferencias no resultaron significativas (p-valor=0,09). Por otro lado la cantidad de *InDel* disminuyó, siendo esta reducción significativa (p-valor<0,001). Esta tendencia general se mantiene en la mayoría de los cromosomas, con excepción de los cromosomas 7 y 8 donde se redujo el número de *SNP* detectado en el alineado a la versión SL3.0 del genoma consenso; y en el cromosoma 10 donde se incrementó la cantidad de *InDel* detectada (Tabla 1, Figura 3). La menor cantidad de *InDel* en la versión SL3.0, podría deberse a una disminución de la cantidad de espacios de secuencia desconocida (*GAP*) lo cual puede dar lugar a una reducción en la detección de falsos *InDel* y la identificación de nuevos *SNP*. Por su parte se advierte que la cantidad de *SNP* fue mayor que la cantidad de *InDel* en todos los cromosomas y en ambas versiones del genoma. Las mutaciones puntuales son más abundantes en la naturaleza que aquellas que implican cambio en más de una base, por lo cual este resultado es esperable.

Figura 3: Cantidad de polimorfismos tipo *SNP* e *InDel* identificados en ambos *bulks* respecto a diferentes versiones (SL2.50 y SL3.0) del genoma de referencia.



La Figura 4 presenta la cantidad de *SNP* y de *InDel* detectados en función de la profundidad de cobertura para ambas versiones del genoma de referencia. La cantidad detectada de ambos tipos de polimorfismos fue semejante en función de la profundidad de cobertura para ambas versiones del genoma de referencia. Se observa que la profundidad de lectura media fue de 50 pb para los *SNP* y levemente menor para los *InDel.*

Figura 4: Cantidad de *SNP* e *InDel* en función de la profundidad de cobertura, identificados en ambos *bulks* respecto a diferentes versiones (SL2.50 y SL3.0) del genoma de referencia. a) y b) Cantidad de *SNP* identificados respecto de la versión SL2.50 y SL3.0, respectivamente. c) y d) Cantidad de *InDel* identificados respecto de la versión SL2.50 y SL3.0, respectivamente.



Se analizó la cantidad total de *InDel* de distintas longitudes, en cada cromosoma y en las dos versiones de la referencia. Los resultados se resumen en la Tabla 2. Se observó en ambos alineamientos, que los *InDel* más abundantes fueron aquellos de longitud menor a 5 pb. La cantidad de *InDel* disminuyó cuanto más largo es el polimorfismo, exceptuando aquellos que superan los 50 pb que fueron más numerosos. Esto se puede deber a que esta categoría incluye los errores de alineado, siendo la mayoría probablemente falsos *InDel*. Por su parte, es remarcable que los resultados en los alineados realizados para ambas versiones del genoma son muy similares.

Uno de los principales objetivos del mejoramiento vegetal es identificar regiones genómicas asociadas a caracteres de interés e introducirlos a cultivares elite mediante el método de retrocruzas. El uso de marcadores moleculares estrechamente ligados al carácter de interés permite determinar en estadíos tempranos de desarrollo aquellos individuos que llevan las regiones genómicas deseadas. A esta estrategia se la denomina selección asistida por marcadores (MAS, por sus siglas en inglés *Marker-assisted Selection*) (Ashikari y Matsuoka, 2006) y permite reducir el trabajo y tiempos requeridos para la evaluación fenotípica y mejorar la tasa de introgresión (Collard y Mackill, 2008; Ribaut y Hoisington, 1998). Existe una gran variedad de marcadores y metodología disponibles para identificar las diferencias genómicas entre los individuos. Los *InDel* constituyen la segunda forma de variación en el genoma, y pueden generar efectos fenotípicos importantes. Debido a que la

técnica y los equipamientos demandados resultan relativamente simples y poco costosos, el desarrollo y genotipado con marcadores de ADN de tipo *InDel* constituye una estrategia accesible en los programas de manejo que no tienen acceso al genotipado de alta resolución basado en marcadores tipo *SNP* (Cambiaso et al., 2019). Por lo general estos marcadores incluyen un par de cebadores en sitios únicos flanqueantes a los *InDel* cuyo tamaño va de 1 a 100 pb. Estas regiones se amplifican mediante una reacción en cadena de la polimerasa (PCR, del inglés *Polymerase Chain Reaction*) y los amplicones son visualizados en geles de agarosa o poliacrilamida.

En función de estos antecedentes, se separaron los polimorfismos de tipo *InDel* hallados en subgrupos de acuerdo a su tamaño, ya que esto determina la técnica molecular de revelado que se pueden utilizar. Así aquellos *InDel* de tamaño entre 4 y 10 pb solo podrán visualizarse mediante técnicas de revelado de alta resolución, como son los geles de poliacrilamida. Por el contrario, cuando los polimorfismos son de una longitud mayor a 15 pb las diferencias genotípicas pueden visualizarse utilizando geles de agarosa. Se ha probado exitosamente en tomate que se pueden incluir varias combinaciones de cebadores de distinto tamaño en una misma *PCR*, lo que permite ahorrar recursos económicos y tiempo (Cambiaso et al., 2019). En este sentido, resulta de gran interés distinguir *InDel* de distintos tamaños: entre 10 y 20 pb (tamaño chico), de 21 a 35 pb (tamaño mediano) y mayores a 36 pb (tamaño grande) para explotar dicho beneficio.

Tabla 2: Cantidad de *InDel* por cromosoma (cr) en función de su longitud en pares de base (pb).

	SL2.50						SL3.0					
	<5pb	**5-15pb**	**15-20pb**	**21-35pb**	**36-50pb**	**>50**	**<5pb**	**5-15pb**	**15-20pb**	**21-35pb**	**36-50pb**	**>50**
Cr01	14.780	856	131	166	86	120	12.915	847	132	165	86	118
Cr02	8.386	770	110	149	74	73	7.637	790	111	155	82	77
Cr03	11.960	1.111	187	194	74	110	10.681	1.099	184	206	80	107
Cr04	13.055	1.313	188	215	94	125	11.870	1.306	195	213	94	126
Cr05	10.063	936	143	156	77	79	9.064	908	145	153	78	81
Cr06	8.359	590	84	106	47	63	7.356	583	85	117	45	63
Cr07	8.529	538	90	103	60	71	7.334	534	94	102	58	77
Cr08	7.403	639	104	125	57	82	6.621	619	105	114	55	82
Cr09	9.543	726	91	140	68	58	8.703	731	100	132	65	57
Cr10	11.700	943	191	168	85	120	11.734	988	210	189	89	133
Cr11	11.725	1.135	203	195	85	120	10.720	1.129	201	192	86	117
Cr12	9.695	753	121	136	65	98	8.604	732	127	135	69	102
TOTAL	**125.198**	**10.310**	**1.643**	**1.853**	**872**	**1.119**	**113.239**	**10.266**	**1.689**	**1.873**	**887**	**1.140**

Identificación de regiones genómicas asociadas al carácter tipo de carpelo:

La identificación de regiones genómicas asociadas al carácter tipo de carpelo se llevó a cabo a partir de los datos de *SNP* únicamente, ya que son los polimorfismos más abundantes en el genoma. Debido a que los resultados obtenidos en la cantidad y distribución de estos polimorfismos en ambas versiones del genoma, no presentan diferencias significativas entre sí. Con el fin de evitar información redundante, se expresarán solo aquellos resultados obtenidos mediante el alineado a la versión SL3.0 del genoma de tomate, por ser la versión más reciente.

Se identificaron 207.674 *SNP* en ambos grupos respecto a la secuencia consenso. La cobertura promedio fue de 35,23 X en el grupo no fusionado y 27,59 X para el fusionado. Estos valores resultan altos considerando la profundidad de cobertura de otros trabajos en diversas especies (Kadambari et al., 2018 en arroz; Li et al., 2018 en maíz; Zhang et al., 2018 en soja) y también en tomate (Cambiaso et al., 2019; Ruangrak et al., 2018; Wang et al., 2018). Se encontraron *SNP* distribuidos a lo largo de todo el genoma y el porcentaje de cobertura del genoma considerando una profundidad de 4X fue de 89,98% en el grupo no fusionado y 89,91% en el fusionado.

Mediante el filtrado de los datos se eliminaron 74.023 *SNP* que presentaron frecuencia del alelo de referencia menor a 0,2 o mayor a 0,8 en ambos *bulks* ya que podrían deberse a problemas de secuenciación o errores de alineado. También se eliminaron 24.871 *SNP* que tuvieron profundidades de lectura en ambos *bulks* menor a 38 pb y 29.519 *SNP* con profundidad de lectura en ambos *bulks* mayor a 65 pb. Además, se removieron 13 *SNP* que presentaron una profundidad de lectura individual menor a 5 pb y 11.854 *SNP* con calidad inferior a 99% (probabilidad de error mayor al 1%). De este modo de los 207.674 *SNP* detectados inicialmente, se filtraron 140.280 y se conservaron 67.394 (32,45%) que fueron considerados en el análisis. La aplicación de filtros rigurosos aumenta la confiabilidad de los resultados obtenidos. Filtrando en base a la profundidad de lectura se busca eliminar *SNP* con extremadamente baja o alta cobertura, debido a que una baja cobertura implica baja confiabilidad, mientras que valores muy altos podrían indicar zonas repetitivas, lo que conduce a una sobreestimación de la profundidad de lectura. Los datos de *SNP* se sometieron a un análisis de grupos segregantes basados en secuenciación luego del filtrado.

Como era esperable, los valores de índice-*SNP* se distribuyeron principalmente en forma aleatoria alrededor del valor de 0,5 para la mayoría de las partes del genoma en ambos *bulks*. La dispersión de los valores de índice-*SNP* entorno al valor de 0,5 se debe a la gran cantidad de *SNP* y la naturaleza estocástica de la segregación alélica en cada SNP individual, que sigue una distribución binomial con un parámetro de probabilidad de 0,5 (Abe et al., 2012). La presencia de una región genómica única con un grupo de *SNP* con valor de índice-*SNP* de uno, indica la posible ubicación de la mutación responsable de un fenotipo mutante. Sin embargo, no es necesario identificar la mutación causal del genotipo sino que es posible utilizar los *SNP* que

flanquean las regiones que albergan mutaciones causales como marcadores de ADN para la selección asistida por marcadores durante el mejoramiento del cultivo (Abe et al., 2012). En el presente trabajo no fue posible identificar valores de índice-*SNP* de 1, siendo los máximos valores encontrados 0,92 en el grupo no fusionado y 0,93 en el grupo fusionado. Esto se debe a que el análisis se centra en un carácter cualitativo, donde el fenotipo fusionado resulta dominante sobre el no fusionado. En consecuencia, el efecto de alelo recesivo se encuentra encubierto por el alelo dominante, siendo imposible diferenciar los individuos homocigotas dominantes (presentan únicamente la variante alélica dominante para un carácter) de aquellos heterocigotas (tienen la variante alélica dominante y la variante recesiva para el carácter). Así el grupo fusionado presentará una mezcla de individuos homocigotas y heterocigotas, mientras que el otro grupo incluirá solo individuos homocigotas como el fenotipo recesivo. Por este motivo no existirá una dispersión clara de los valores fenotípicos disminuyendo el valor absoluto de Δ(índice-*SNP*).

Considerando los valores de G', se detectaron tres regiones genómicas asociadas al carácter tipo de carpelo ubicadas en los cromosomas 3, 6 y 10. La región del cromosoma 3 presentó una longitud de 2,88 Mb, 250 *SNP* polimórficos entre los *bulks*, un valor máximo de Δ(índice-*SNP*)= -0,25 y de G' = 4,69 en las posiciones 56,48 Mb. El valor de Q promedio en la región fue 0,03. El *QTL* ubicado en el cromosoma 6 tiene una longitud de 9 Mb, un valor máximo de Δ(índice-*SNP*)= 0,31 y G'= 9,8 en las posiciones 45,11 y 43,57 Mb, respectivamente. El valor de Q promedio en la región fue de 0,017 y presentó 1.137 *SNP* polimórficos entre los *bulks*. La región ubicada en el cromosoma 10 fue la más extensa, con una longitud de 59 Mb, y presentó un máximo valor de Δ(índice-*SNP*)= 0,32 y G'=7,27 en las posiciones 44,24 Mb y 27,95 Mb respectivamente. Por su parte esta región mostró un valor de Q promedio de 0,006 y contuvo 8.097 *SNP* polimórficos entre los *bulks*.

Los picos o máximos valores de G' que superan los valores de FDR (q) de 0,05 se consideraron significativos, es decir que dichas regiones probablemente contengan genes relacionados al carácter tipo de carpelos. Estos picos indican que existen frecuencias alélicas diferenciales significativas entre los grupos, es decir que un grupo presenta una sobre o sub representación del genoma de un parental mientras que en el otro grupo sucede lo opuesto. Esto sugiere que dichas regiones presentan un efecto significativo sobre el carácter. Los umbrales de significancia fueron estimados a partir de la distribución nula de G', suponiendo la ausencia de *QTL* vinculados al *SNP* de interés, por lo tanto aquellos valores de G' que superan los umbrales son considerados raros y constituyen los *QTL* putativos (Magwene et al., 2011; Yang et al., 2013).

Dado que el máximo valor de G' se encontró en el cromosoma 6, nos permite postular a dicha región como la principal candidata a controlar el carácter tipo de carpelo.

La direccionalidad del Δ(índice-*SNP*) es también relevante. Un valor de Δ (índice-*SNP*) mayor a cero indica que el progenitor que contribuye es aquel que tiene el alelo alternativo, mientras que un valor de Δ(índice-*SNP*) menor a cero o negativo indica que el padre contribuyente al carácter es aquel que tiene el alelo como la referencia. En nuestros datos se observa que en la región subyacente al carácter presente en el cromosoma 3 los alelos son aportados por el padre como la referencia (*S. lycopersicum* L. cv. 'Old Brooks'), mientras que en los *QTL* ubicados en los cromosomas 6 y 10 los alelos provienen del otro progenitor que presenta la otra variante (*S. lycopersicum* L. cv. 'Voyage').

Tabla 3: Descripción de largo, cantidad de *SNP*, valor y posición de Δ(índice-*SNP*) y G´ máximos y significancia hallados para los *QTL* putativos identificados en los distintos cromosomas (versión del genoma SL 3.0) para el carácter tipo de carpelos

Cromosoma	Comienzo	Fin	Largo	n*SNP*	pico Δ(índice-*SNP*)	Posición pico	Máximo G´	Posición max.G´	p-value medio	Q-value medio
SL3.0cr03	55190911	58066873	2875962	250	-0,249	56482547	4,687	56482547	0,004	0,030
SL3.0cr06	40331359	49361242	9029883	1137	0,308	45114796	9,82	43569191	0,002	0,017
L3.0cr10	2031647	61900727	59869080	8097	0,324	44241355	7,267	27947861	0,001	0,006

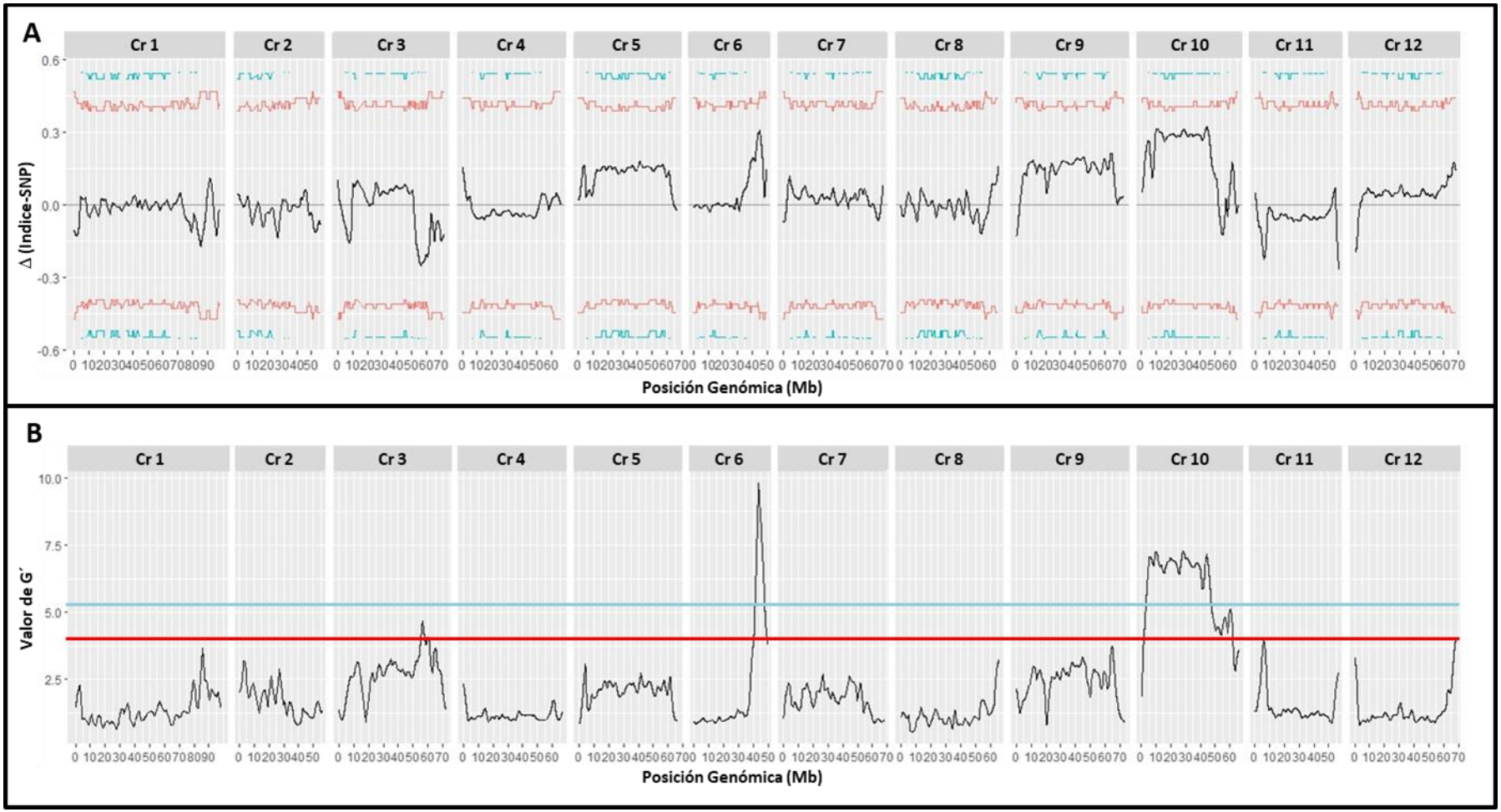
Comienzo: posición genómica de inicio. Fin: posición final del *QTL*. Largo: longitud en bases desde el comienzo al fin de la región. n*SNP*: cantidad de *SNP* en la región. pico Δ(índice-*SNP*): valor de Δ(índice-*SNP*) en el pico. Posición pico: posición del valor absoluto de Δ(índice-*SNP*) suavizado por tricubo máximo. Máximo G´: valor máximo de G´ en la región. Posición máx. G´: posición genómica del máximo valor de G´en el *QTL*. p-valor medio: valor de p promedio en la región.

Esquematización de *QTL* putativos:

En la Figura 5 A se observan regiones con un valor absoluto de Δ(índice-*SNP*) mayor a cero. Esto revela que existe una representación desigual del genoma de los progenitores. A lo largo de la mayor parte del genoma las variaciones son al azar y los valores de Δ(índice-*SNP*) son bajos, sin embargo se pueden distinguir un valle más pronunciado en la base del cromosoma 3 y dos picos, uno en la base del cromosoma 6 y otro a lo largo del cromosoma 10. Esto sugiere que dichas regiones estarían asociadas a las diferencias fenotípicas entre los grupos, en este caso tipo de carpelo fusionado o no fusionado. A pesar de ello en ningún caso el valor absoluto Δ (índice-*SNP*) superó al límite del intervalo de confianza. Por lo tanto, estas regiones no difieren significativamente de cero y no hay evidencia en los datos para confirmar que existe asociación de las mismas con el carácter tipo de carpelos.

Al analizar los valores de G´ (Figura 5 B) se pueden identificar regiones en los cromosomas 3, 6 y 10 cuyo valor de G´ sobrepasan el umbral de significancia FRD (q) de 0,05, coincidentes con aquellas regiones de valles y picos pronunciados aunque no significativos, detectados en la Figura 5 A. Si analizamos el umbral FRD (q) de 0,01 solo los picos presentes en los cromosomas 6 y 10 resultan significativos. Esto implica que existen regiones genómicas putativas asociadas al tipo de carpelo en los cromosomas 6 y 10 con un 99% de significancia y en el cromosoma 3 con un 95% de significancia. También se evidencia que el pico más significativo estuvo presente en el cromosoma 6.

Figura 5: Regiones genómicas asociadas al carácter tipo de carpelo a lo largo de los 12 cromosomas de tomate, identificados por *QTL*seqr.



A) Valores de∆(índice-*SNP*) suavizado por tricubo. Intervalo de confianza de 95% indicado en rojo y de 99% indicado en azul. B) Valores de G´ suavizada por tricubo. Valor de significancia de FDR (Q): 0,05 indicado en rojo y valor de FDR (Q): 0,01 indicado en azul. Ventana deslizante=1Mb.

Según lo descripto por Takagi et al. (2013) la detección de *QTL* mediante *QTL-seq* depende de distintas variables experimentales que afectan el desempeño de la metodología. Mediante un proceso de simulación computacional del análisis de *QTL-seq* modificando distintas variables fue posible identificar los parámetros que determinan una mayor potencia de detección de *QTL*, entendiendo como potencia a la proporción de las repeticiones de simulación con una significancia mayor que el valor de corte del 99%. Los resultados mostraron que la potencia de detección de esta metodología aumentó con mayor profundidad de cobertura, mayor porcentaje de individuos considerados en los *bulks*, mayor contribución del *QTL* a la variación fenotípica (heredabilidad) y a su vez cuando el efecto de los alelos es codominante. Magwene et al. (2011) por su parte plantea resultados similares. Se observó que existe un valor óptimo de porcentaje de individuos seleccionados para los *bulk* (Magwene et al., 2011; Takagi et al., 2013; Wang et al., 2019) ya que cuando el número de individuos seleccionados es reducido, la potencia de detección es baja, probablemente por el alto efecto de muestreo, luego el poder aumenta a medida que aumenta el porcentaje, hasta un punto donde comienza a disminuir, cuando el tamaño es tan grande que se comienza a incluir individuos con fenotipos intermedios en los *bulks*. En otras palabras, *bulks* muy grandes implican una selección más débil y, por lo tanto, una divergencia de frecuencias alélicas más pequeña entre los *bulks*. Sin embargo al disminuir el tamaño de *bulk* aumenta la intensidad de selección y la divergencia de las frecuencias alélicas entre los *bulks*, pero en menor proporción que el aumento de la varianza de las frecuencias alélicas en cada *bulk*. Por su parte, la potencia es mayor en las poblaciones tipo RIL que en las F_2, sin embargo con una profundidad de cobertura y porcentaje de individuos seleccionados adecuados es posible igualar la potencia de detección alcanzada en las poblaciones RIL. La principal ventaja de utilizar una población de mapeo del tipo F_2 es el corto tiempo necesario para su obtención, a pesar de que las mismas no son replicables. El tamaño de ventana deslizante en el análisis es otro factor de consideración al realizar el proceso de suavizado por tricubo. Así a mayor tamaño de la ventana, menor es la frecuencia de las señales que son eliminadas por el filtro. Por lo tanto la elección del ancho de suavizado implica un compromiso entre filtrar las desviaciones de alta frecuencia en G y atenuar la señal de *QTL* reales (Magwene et al., 2011). Finalmente se puede postular que el tamaño de población afectará la detección de regiones genómicas significativas al afectar la determinación por simulación de los intervalos de confianza. Cuando se simulan las situaciones extremas de *QTL* para calcular los intervalos de confianza, un tamaño pequeño de población aumenta la probabilidad de obtener por azar un valor alto de Δ(índice-*SNP*). En consecuencia a menor tamaño de población, menor la confianza de los datos y mayor intervalo (Mansfeld y Grumet, 2018).

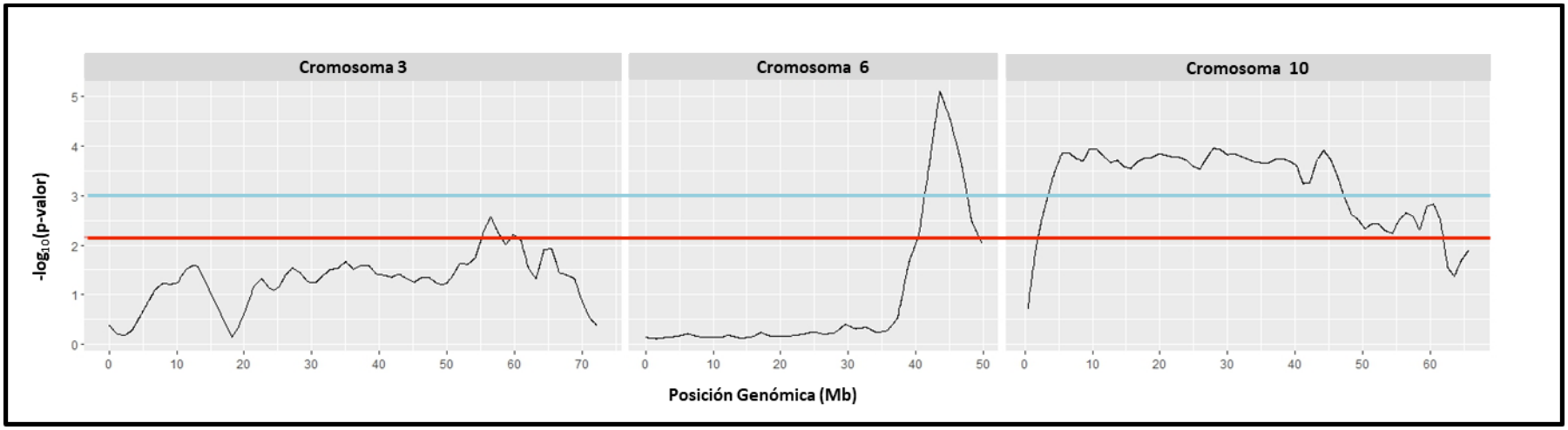
En función de estas consideraciones podemos hipotetizar que la baja capacidad de detección de regiones significativas mediante la metodología de *QTL-seq* en nuestro estudio se debe al reducido tamaño de población (76 individuos), lo que disminuye la

confianza de los valores de Δ(índice-*SNP*). Por su parte dada la naturaleza cualitativa del carácter tipo de carpelo, donde el fenotipo no fusionado es recesivo, no permite la dispersión clara de los datos fenotípicos y es esperable que el *bulk* fusionado incluya una mezcla de individuos con genotipo homocigota y heterocigota; mientras que el otro *bulk* presentará solo individuos homocigotas como el recesivo. Esta modificación en la conformación de los *bulks* determina que el valor absoluto de Δ(índice-*SNP*) disminuya. Sin embargo, debido a estas consideraciones, se puede decir que el carácter tipo de carpelo estaría poco afectado por los factores ambientales, presentando alta heredabilidad y en consecuencia se deben considerar los tres picos más significativos como potenciales regiones de interés.

El uso de la metodología basada en la estadística G proporciona varias ventajas sobre el uso de las diferencias de frecuencia de los alelos como base para la estimación de *QTL* (por ejemplo, Parts et al., 2011). Se espera que G disminuya mucho más rápidamente alrededor del sitio de interés que el sesgo en las frecuencias alélicas, lo que implica intervalos de apoyo más estrechos alrededor de los *QTL*. También en contraste con las estadísticas basadas en la divergencia de las frecuencias alélicas, G tiene en cuenta la solidez de la evidencia relacionada con el tamaño de la muestra. Esta característica del estadístico G también puede complicar potencialmente los análisis, ya que la variación en la profundidad de lectura contribuye a la variación en G en escalas espaciales relativamente pequeñas. Sin embargo, el promedio ponderado de G suaviza efectivamente la variabilidad de la "alta frecuencia" asociada con la variación de secuencia (Magwene et al., 2011). El efecto del promediado de G', reduce el efecto de regiones sesgadas hacia nucleótidos particulares durante el proceso de secuenciación, lecturas mal ajustadas o *SNP* erróneos. Sin embargo, hay que considerar las regiones genómicas que son particularmente problemáticas, como las ricas en regiones repetitivas. Usando la estadística G suavizada, o G' es posible reducir la dispersión de los datos al mismo tiempo que se aborda el desequilibrio de ligamiento entre los *SNP*. Además, como G' está cerca de ser log normal, los valores de p se pueden estimar para cada *SNP* usando una estimación no paramétrica de la distribución nula de G'. Esto proporciona un resultado claro y fácil de interpretar, así como la opción para múltiples correcciones de prueba (Mansfeld y Grumet, 2018).

La visualización de los resultados como el $-\log_{10}$ del valor de p derivado del valor de G´ (Figura 6) resulta una forma más intuitiva y sencilla para identificar las regiones significativas asociadas al carácter de interés y distinguir con más detalle el inicio y fin de dichas regiones. Como se puede observar en la Figura 6, las regiones genómicas putativas asociadas a tipo de carpelo se ubican en la parte basal del cromosoma 6, la región centromérica del cromosoma 10 y la parte basal del cromosoma 3. Por su parte la región del cromosoma 10 abarca prácticamente todo el cromosoma.

Figura 6: *QTL* putativos para el carácter tipo de carpelo.



Valores de $-\log_{10}$ (p-valor) derivado del valor de G´. La línea sólida de color rojo y azul indican los valores de significancia FDR (Q) 0,05 y 0,01 respectivamente. Ventana deslizante=1Mb

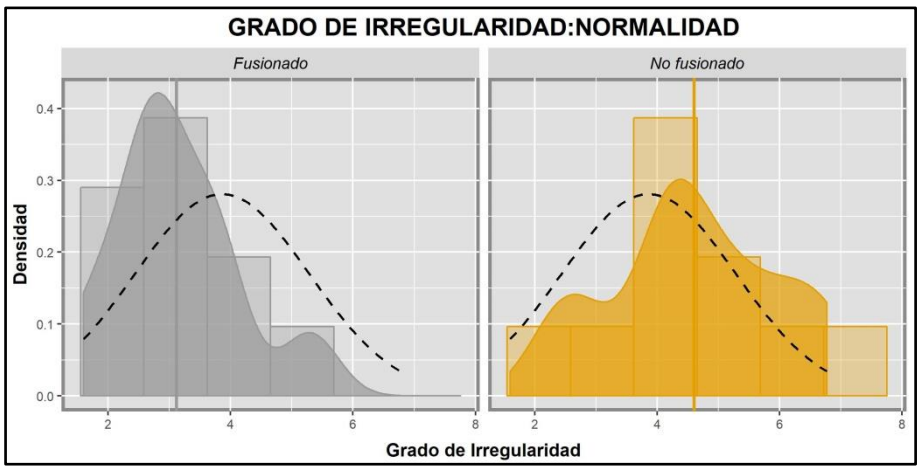
.

Comparación de datos fenotípicos de los grupos segregantes para un carácter cuantitativo:

En este trabajo se desarrolló la técnica de *QTL-seq* para detectar regiones genómicas de tomate asociadas al carácter discreto tipo de carpelos. Sin embargo es de interés conocer si este carácter se corresponde con el carácter cuantitativo grado de irregularidad externa.

Se observó que el carácter Grado de Irregularidad Externa de los frutos se ajusta a una distribución normal en ambos *bulks* (Fusionados y No fusionados) (Figura 7).

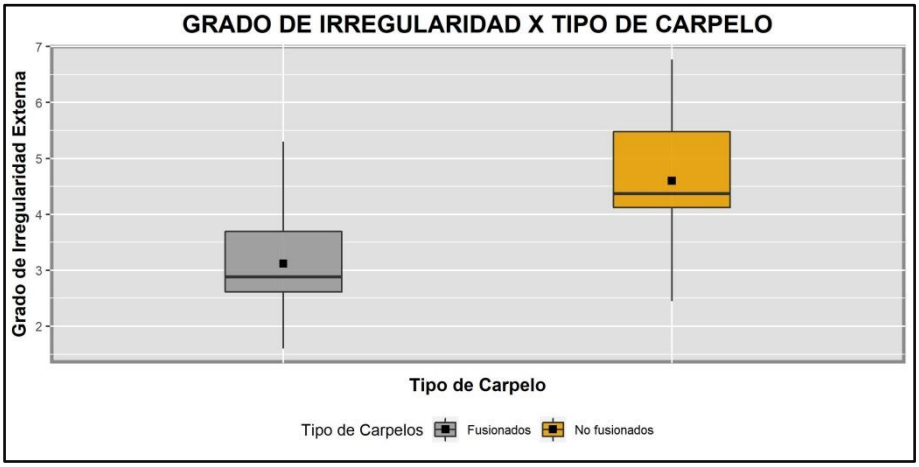
Figura 7: Análisis de normalidad del carácter Grado de Irregularidad Externa para ambos bulks



Barras muestran el histograma de densidad para el carácter grado de irregularidad externa discriminado por tipo de carpelo (gris= fusionado, amarillo= no fusionado) y las curvas la densidad. Líneas punteadas negras indican la distribución normal.

Existen diferencias significativas entre ambos *bulks* para grado de irregularidad externa (t=-2,68 y p-valor<0,05). Es decir que aquellas plantas que tienen fenotipo no fusionado para el carácter tipo de carpelo también presentan alto grado de irregularidad, mientras que aquellas plantas con frutos de carpelos fusionados tienen a su vez forma más regular (Tabla 4, Figura 8). Esto indicaría que las mismas regiones genómicas que controlan el carácter tipo de carpelo estarían controlando el carácter grado de irregularidad externa de los frutos. Es decir que se trataría del mismo gen con un efecto pleiotrópico sobre el carácter cuantitativo o dos genes ligados en la misma región. En este sentido es posible, asociar un carácter cualitativo a otro cuantitativo, siendo correcta la metodología implementada.

Figura 8: Comparación de medias del carácter Grado de Irregularidad Externa, en función de tipo de carpelos.



Valores máximos, mínimo, mediana, cuartil 1 y cuartil 3, para el carácter grado de Irregularidad Externa en función de tipo de carpelos. Los cuadrados negros muestran el valor medio.

Tabla 4: Medidas descriptivas de ambos grupos (fusionado y no fusionado) para el carácter grado de irregularidad externa.

Grupo	Media	D.E.	Mín	Máx	Mediana	Q1	Q3
Fusionado	31,20	10,34	16,00	53,00	28,85	25,70	37,90
No fusionado	46,02	14,03	24,50	67,70	43,70	40,90	55,90

D.E.: desvío estándar, Máx: valor máximo, Mín: valor mínimo, Q1: cuartil 1 y Q3: cuartil 3, para el carácter grado de Irregularidad Externa en función de tipo de carpelos.

CONCLUSIÓN

Se alinearon las secuencias genómicas completas de grupos discrepantes para el carácter tipo de carpelos en tomate a la secuencia de referencia de tomate.

La comparación entre ambos grupos permitió identificar tres regiones genómicas con polimorfismos diferenciales significativos asociados al carácter de interés, tipo de carpelo en tomate.

Se plantea para un futuro, en un principio validar las regiones putativas detectadas mediante marcadores moleculares en poblaciones segregantes derivadas del cruzamiento entre los cultivares de tomate *(S. lycopersicum L.)* 'Voyage' y 'Old Brooks'. También se definirá con mayor exactitud la ubicación de las regiones asociadas al carácter de interés. Una vez detectados marcadores moleculares altamente asociados al carácter, se podrán utilizar los mismos en programas de mejoramiento del cultivo de tomate para realizar *MAS* y en otros trabajos de investigación. Finalmente se realizará un estudio de los genes candidatos presentes en dichas regiones para identificar el gen o los genes implicados en la determinación del carácter.

BIBLIOGRAFÍA

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., Matsumura, H., Yoshida, K., Mitsuoka, C., Tamiru, M., Innan, H., Cano, L., Kamoun, S., & Terauchi, R. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature Biotechnology*, *30*(2), 174–178. https://doi.org/10.1038/nbt.2095

Aflitos, S., Schijlen, E., Jong, H., Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L., Bakker, F., Dirks, R., Breit, T., Gravendeel, B., Huits, H., Struss, D., Swanson-Wagner, R., Leeuwen, H., Ham, R. C., … Peters, S. (2014). Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *The Plant Journal*, *80*, 136–148. https://doi.org/10.1111/tpj.12616

Ashikari, M., & Matsuoka, M. (2006). Identification, isolation and pyramiding of quantitative trait loci for rice breeding. *Trends in Plant Science*, *11*(7), 344–350. https://doi.org/https://doi.org/10.1016/j.tplants.2006.05.008

Asnaghi, C., Roques, D., Ruffel, S., Kaye, C., J.-Y. Hoarau, Télismart, H., Girard, J. C., Raboin, L. M., Risterucci, A. M., Grivet, L., & D'Hont, A. (2004). Targeted mapping of a sugarcane rust resistance gene (Bru1) using bulked segregant analysis and AFLP markers. *Theoretical and Applied Genetics*, *108*(4), 759–764.

Auwera, G. A. Van der, Carneiro, M. O., Hartl, C., Poplin, R., Angel, G. del, Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, *43*, 11.10.1–11.10.33. https://doi.org/https://doi.org/10.1002/0471250953.bi1110s43

Babraham Bioinformatics Group. (2017). *Trim galore.*

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Bernatzky, R., & Tanksley, S. D. (1986). Genetics of actin-related sequences in tomato. *Theoretical and Applied Genetics*, *72*, 314–321.

Cambiaso, V., Pratta, G. R., Pereira da Costa, J. H., Zorzoli, R., Francis, D. M., & Rodríguez, G. R. (2019). Whole genome re-sequencing analysis of two tomato genotypes for polymorphism insight in cloned genes and a genetic map construction. *Scientia Horticulturae*, *247*. https://doi.org/10.1016/j.scienta.2018.12.001

Causse, M., Desplat, N., Pascual, L., Le Paslier, M. C., Sauvage, C., Bauchet, G., Bérard, A., Bounon, R., Tchoumakov, M., Brunel, D., & Bouchet, J. P. (2013). Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-791

Chagué, V., Mercier, J. C., Guénard, M., Courcel, A. de, & Vedel, F. (1997). Identification of RAPD markers linked to a locus involved in quantitative resistance to TYLCV in tomato by bulked segregant analysis. *Theoretical and Applied Genetics*, *95*(4), 671–677.

Chen, Q., Song, J., Du, W.-P., Xu, L.-Y., Jiang, Y., Zhang, J., Xiang, X.-L., & Yu, G.-R. (2017).

Identification, mapping, and molecular marker development for Rgsr8.1: a new quantitative trait locus conferring resistance to gibberella stalk rot in maize (Zea mays L.). *Frontiers in Plant Science*, *8*, 1355. https://doi.org/10.3389/fpls.2017.01355

Collard, B. C., & Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1491), 557–572. https://doi.org/https://doi.org/10.1098/rstb.2007.2170

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, *142*(1–2), 169–196. https://doi.org/10.1007/s10681-005-1681-5

Cong, B., Barrero, L. S., & Tanksley, S. D. (2008). Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nature Genetics*, *40*(6), 800–804. https://doi.org/10.1038/ng.144

Das, S., Upadhyaya, H. D., Bajaj, D., Kujur, A., Badoni, S., Laxmi, V. K., Shailesh Tripathi, C. L. L. G., Sharma, S., Singh, S., Tyagi, A. K., & Parida, S. K. (2015). Deploying QTL-seq for rapid delineation of a potential candidate gene underlying major trait-associated QTL in chickpea. *DNA Research*, *22*(3), 193–203. https://doi.org/https://doi.org/10.1093/dnares/dsv004

DePristo, M. A., Rivas, M. A., McKenna, A., Hartl, C., del Angel, G., Sivachenko, A. Y., Philippakis, A. A., Hanna, M., Daly, M. J., Altshuler, D., Gabriel, S. B., Fennell, T. J., Poplin, R., Garimella, K. V, Kernytsky, A. M., Cibulskis, K., Maguire, J. R., & Banks, E. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Falconer, D., & Mackay, T. (1996). Introduction to quantitative genetics. In *London: Prentice Hall: Vol. 4th edn*.

Fernandez-pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., Bombarely, A., Fisher-york, T., Pujar, A., Foerster, H., Yan, A., & Mueller, L. A. (2015). *The Sol Genomics Network (SGN)—from genotype to phenotype to breeding*. *43*(November 2014), 1036–1041. https://doi.org/10.1093/nar/gku1195

Fulton, T. M., Chunwongse, J., & Tanksley, S. D. (1995). Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Molecular Biology Reporter*, *13*(3), 207–209.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., van der Knaap, E., Huang, S., Klee, H. J., … Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, *51*(6), 1044–1051. https://doi.org/10.1038/s41588-019-0410-2

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*, *28*(20), 2678–2679. https://doi.org/10.1093/bioinformatics/bts503

Giovannoni, J. J., Wing, R. A., Ganal, M. W., & Tanksley, S. D. (1991). Isolation of molecular

markers from specific chromosomal intervals using DNA pools from existing mapping populations. *Nucleic Acids Research*, *19*(23), 6553–6568. https://doi.org/10.1093/nar/19.23.6553

Halldén, C., Säll, T., Olsson, K., Nilsson, N. -O., & Hjerdin, A. (1997). The use of bulked segregant analysis to accumulate RAPD markers near a locus for beet cyst nematode resistance in Beta vulgaris. *Plant Breeding*, *116*(1), 18–22. https://doi.org/10.1111/j.1439-0523.1997.tb00970.x

Hampel, F. (1978). Modern trends in the theory of robustness. *Series Statistics: A Journal of Theoretical and Applied Statistics*, *9*(3), 425–442. https://doi.org/0.1080/02331887808801443

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., Dong, G., Sang, T., & Han, B. (2009). High-throughput genotyping by whole-genome resequencing. *Genome Research*, *19*(6), 1068–1076. https://doi.org/10.1101/gr.089516.108

Illa-Berenguer, E., Van Houten, J., Huang, Z., & van der Knaap, E. (2015). Rapid and reliable identification of tomato fruit weight and locule number loci by QTL-seq. *Theoretical and Applied Genetics*, *128*(7), 1329–1342. https://doi.org/10.1007/s00122-015-2509-x

Jung, Y. J., Nou, I. S., Cho, Y. G., Kim, M. K., Kim, H. T., & Kang, K. K. (2016). Identification of an SNP variation of elite tomato (Solanum lycopersicum L.) lines using genome resequencing analysis. *Horticulture Environment and Biotechnology*, *57*(2), 173–181. https://doi.org/10.1007/s13580-016-0132-7

Kadambari, G., Vemireddy, L. R., & Srividhya, A. (2018). QTL-Seq-based genetic analysis identifies a major genomic region governing dwarfness in rice (Oryza sativa L .). *Plant Cell Reports*, *37*(4), 677–687. https://doi.org/10.1007/s00299-018-2260-2

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, P., Du, C., Zhang, Y., Yin, S., Zhang, E., Fang, H., Lin, D., Xu, C., & Yang, Z. (2018). Combined bulked segregant sequencing and traditional linkage analysis for identification of candidate gene for purple leaf sheath in maize. *PLOS ONE*, *13*, 1–14. https://doi.org/10.1371/journal.pone.0190670

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., Huang, Z., Li, J., Zhang, C., Wang, T., Zhang, Y., Wang, A., Zhang, Y., Lin, K., Li, C., … Huang, S. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics*, *46*(11), 1220–1226. https://doi.org/10.1038/ng.3117

Lippman, Z., & Tanksley, S. D. (2001). *Dissecting the Genetic Pathway to Extreme Fruit Size in Tomato Using a Cross Between the Small-Fruited Wild Species Lycopersicon pimpinellifolium and*. *1999*.

Liu, J., Van Eck, J., Cong, B., & Tanksley, S. D. (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proceedings of the National Academy of Sciences*, *99*(20), 13302–13306. https://doi.org/10.1073/pnas.162485999

Lu, H., Lin, T., Klein, J., Wang, S., Qi, J., Zhou, Q., Sun, J., Zhang, Z., Weng, Y., & Huang, S. (2014). QTL-seq identifies an early flowering QTL located near Flowering Locus T in cucumber. *Theoretical and Applied Genetics*, *127*(7), 1491–1499. https://doi.org/10.1007/s00122-

014-2313-z

Magwene, P. M., Willis, J. H., & Kelly, J. K. (2011). The statistics of bulk segregant analysis using next generation sequencing. *PLoS Computational Biology*, *7*(11), 1–9. https://doi.org/10.1371/journal.pcbi.1002255

Mansfeld, B. N., & Grumet, R. (2018). QTLseqr: An R package for bulk segregant analysis with next-generation sequencing. *Plant Genome*, *11*(2), 1–5. https://doi.org/10.3835/plantgenome2018.01.0006

Mansur, L. M., OrfK., J., & Lark, G. (1993). Determining the linkage of quantitative trait loci to RFLP markers using extreme phenotypes of recombinant inbreds of soybean (Glycine max L. Merr.). *Theoretical and Applied Genetics*, *86*(8), 914–918.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. a. (2010). The genome analysis toolkit: a mapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Michelmore, R. W., Paran, I., & Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences*, *81*(21), 9828–9832. https://doi.org/10.1073/pnas.88.21.9828

Mueller, L. A., Tanksley, S. D., Giovannoni, J. J., Eck, J. Van, Stack, S., Choi, D., Kim, B. D., Chen, M., Cheng, Z., Li, C., Ling, H., Xue, Y., Seymour, G., Bishop, G., Bryan, G., Sharma, R., & Khurana, J. (2005). *The Tomato Sequencing Project , the first cornerstone of the International Solanaceae Project (SOL)*. *February*, 153–158. https://doi.org/10.1002/cfg.468

Muños, S., Ranc, N., Botton, E., Berard, A., Rolland, S., Duffe, P., Carretero, Y., Le Paslier, M.-C., Delalande, C., Bouzayen, M., Brunel, D., & Causse, M. (2011). Increase in Tomato Locule Number Is Controlled by Two Single-Nucleotide Polymorphisms Located Near WUSCHEL. *Plant Physiology*, *156*(4), 2244–2254. https://doi.org/10.1104/pp.111.173997

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability & Its Applications*, *9*(1), 141–142. https://doi.org/https://doi.org/10.1137/1109020

Parts, L., Cubillos, F. A., Warringer, J., Jain, K., Salinas, F., Bumpstead, S. J., Molin, M., Zia, A., Simpson, J. T., Quail, M. A., Moses, A., Louis, E. J., Durbin, R., & Liti, G. (2011). *Revealing the genetic structure of a trait by sequencing a population under selection*. 1131–1138. https://doi.org/10.1101/gr.116731.110.Freely

Pereira da Costa, J. H., Vega, T. A., Pratta, G. R., Picardi, L. A., Zorzoli, R., & Rodríguez, G. R. (2017). SHORT COMMUNICATION A 54-kDa polypeptide identified by 2D-PAGE and bulked segregant analysis underlies differences for pH values in tomato fruit. *Acta Physiol Plant*, *39*(78). https://doi.org/10.1007/s11738-017-2386-9

Pineda, B., Moreno, V., Lozano, R., Fernández-Lozano, A., Yuste-Lisbona, F. J., Angosto, T., & Pérez-Martín, F. (2014). Mutation at the tomato EXCESSIVE NUMBER OF FLORAL ORGANS (ENO) locus impairs floral meristem development, thus promoting an increased number of floral organs and fruit size. *Plant Science*, *232*, 41–48. https://doi.org/10.1016/j.plantsci.2014.12.007

Quarrie, S. A., Lazić-Jančić, V., Kovačević, D., Steed, A., & Pekić, S. (1999). Bulk segregant

analysis with molecular markers and its use for improving drought resistance in maize. *Journal of Experimental Botany*, *50*(337), 1299–1306.

RCore Team. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Ribaut, J.-M., & Hoisington, D. (1998). Marker-assisted selection: new tools and strategies. *Trends in Plant Science*, *3*(6), 236–239. https://doi.org/DOI:https://doi.org/10.1016/S1360-1385(98)01240-0

Rodríguez, G. R., Kim, H. J., & Van Der Knaap, E. (2013). Mapping of two suppressors of OVATE (sov) loci in tomato. *Heredity*, *111*(3), 256–264. https://doi.org/10.1038/hdy.2013.45

Rodríguez, G. R., Muños, S., Anderson, C., Sim, S.-C., Michel, A., Causse, M., Gardener, B. B. M., Francis, D., & van der Knaap, E. (2011). Distribution of SUN, OVATE, LC, and FAS in the Tomato Germplasm and the Relationship to Fruit Shape Diversity. *Plant Physiology*, *156*(1), 275–285. https://doi.org/10.1104/pp.110.167577

Rodríguez, Gustavo R., Strecker, J., Njanji, I., Thomas, J., Jack, A., Francis, D. M., & van der Knaap, E. (2011). New features and many Improvements to analyze morphology and color of digitalized plant organs are available in Tomato Analyzer 3.0. *Proceedings of the Twenty-Second Midwest Artificial Intelligence and Cognitive Science Conference*, *Vol-710*, 160–163. http://ceur-ws.org/

Ruangrak, E., Su, X., Huang, Z., Wang, X., Guo, Y., Du, Y., & Gao, J. (2018). Fine mapping of a major QTL controlling early flowering in tomato using QTL-seq. *Canadian Journal of Plant Science*, *98*(3), 1–11. https://doi.org/10.1139/cjps-2016-0398

Schucany, W. R. (2004). Kernel Smoothers: An Overview of Curve Estimators for the First Graduate Course in Nonparametric Statistics. *Statistical Science*, *19*(4), 663–675. https://doi.org/https://www.jstor.org/stable/4144437

Shapiro, S. S., & Wilk, M. B. (1965). An analysiAn analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/https://doi.org/10.1093/biomet/52.3-4.591

Song, J., Li, Z., Liu, Z., Guo, Y., & Qiu, L.-J. (2017). Next-Generation Sequencing from Bulked-Segregant Analysis Accelerates the Simultaneous Identification of Two Qualitative Genes in Soybean. *Frontiers in Plant Science*, *8*, 919. https://doi.org/10.3389/fpls.2017.00919

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., & Uemura, A. (2013). *QTL-seq : rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations*. 174–183. https://doi.org/10.1111/tpj.12105

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., Innan, H., Cano, L. M., Kamoun, S., & Terauchi, R. (2013). QTL-seq: Rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant Journal*, *74*(1), 174–183. https://doi.org/10.1111/tpj.12105

Tanksley, S. D. (2009). The Genetic , developmental, and molecular bases of fruit size in tomato and shape variation. *Plant Cell*, *16*(2004), 181–190. https://doi.org/10.1105/tpc.018119.S182

The Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into

fleshy fruit evolution. *Nature*, *485*(7400), 635–641. https://doi.org/10.1038/nature11119

Wang, C., Tang, S., Zhan, Q., Hou, Q., Zhao, Y., Zhao, Q., Feng, Q., Zhou, C., Lyu, D., Cui, L., Li, Y., Miao, J., Zhu, C., Lu, Y., Wang, Y., Wang, Z., Zhu, J., Shangguan, Y., Gong, J., … Han, B. (2019). Dissecting a heterotic gene through GradedPool-Seq mapping informs a rice-improvement strategy. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-11017-y

Wang, R., Sun, L., Bao, L., Zhang, J., Jiang, Y., Yao, J., Song, L., Feng, J., Liu, S., & Liu, Z. (2013). Bulk segregant RNA-seq reveals expression and positional candidate genes and allele-specific expression for disease resistance against enteric septicemia of catfish. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-929

Wang, Y., Jiang, J., Zhao, L., Zhou, R., Yu, W., & Zhao, T. (2018). Application of Whole Genome Resequencing in Mapping of a Tomato Yellow Leaf Curl Virus Resistance Gene. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-27925-w

Watson, G. S. (1964). Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *26*(4), 359–372. https://doi.org/http://www.jstor.org/stable/25049340

Wu, S., Zhang, B., Keyhaninejad, N., Rodríguez, G. R., Kim, H. J., Chakrabarti, M., Illa-Berenguer, E., Taitano, N. K., Gonzalo, M. J., Díaz, A., Pan, Y., Leisner, C. P., Halterman, D., Buell, C. R., Weng, Y., Jansky, S. H., van Eck, H., Willemsen, J., Monforte, A. J., … van der Knaap, E. (2018). A common genetic mechanism underlies morphological diversity in fruits and other plant organs. *Nature Communications*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-07216-8

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., & Van Der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, *319*(5869), 1527–1530. https://doi.org/10.1126/science.1153040

Yang, J., Wang, Y., Shen, H., & Yang, W. (2014). In Silico identification and experimental validation of insertion-deletion polymorphisms in tomato genome. *DNA Research*, *21*(4), 429–438. https://doi.org/10.1093/dnares/dsu008

Yang, Z., Huang, D., Tang, W., Zheng, Y., Liang, K., Cutler, A. J., & Wu, W. (2013). Mapping of Quantitative Trait Loci Underlying Cold Tolerance in Rice Seedlings via High-Throughput Sequencing of Pooled Extremes. *PLoS ONE*, *8*(7). https://doi.org/10.1371/journal.pone.0068433

Zhang, G., Chen, L., Xiao, G., Xiao, Y., Chen, X., & Zhang, S. (2009). Bulked Segregant Analysis to Detect QTL Related to Heat Tolerance in Rice (Oryza sativa L.) Using SSR Markers. *Agricultural Sciences in China*, *8*(4), 482–487. https://doi.org/https://doi.org/10.1016/S1671-2927(08)60235-7

Zhang, X., Wang, W., Guo, N., Zhang, Y., Bu, Y., Zhao, J., & Xing, H. (2018). Combining QTL-seq and linkage mapping to fine map a wild soybean allele characteristic of greater plant height. *BMC Genomics*, *19*(1). https://doi.org/10.1186/s12864-018-4582-4

ANEXO

Detalle de los programas y sentencias aplicados para llevar acabo el alineado de las secuencias genómicas al genoma de referencia, detección de variantes genotípicas o polimorfismos en los *bulks*, comparación de las secuencias y detección de regiones genómicas asociadas al carácter tipo de carpelo.

Alineado de secuencias genómicas:

1. Crear un genoma de referencia con bowtie2

PATH/TO/APP/bowtie2-build -h -s ./PATH/TO/FOLD/REFERENCE_DATA.fa ./PATH/TO/FOLD /REFERENCE_NAME

2. Cortar los adaptadores con trim galore:

/PATH/TO/APP/trim_galore --paired --length 60 --stringency 6 --fastqc --output_dir /PATH/TO/FOLD </PATH/TO/FOLD/INPUT_bulk1_1.fastq.gz> </PATH/TO/FOLD/INPUT_ bulk1_2.fastq.gz>

/PATH/TO/APP/trim_galore --paired --length 60 --stringency 6 --fastqc --output_dir PATH/TO/FOLD </PATH/TO/FOLD/INPUT_bulk2_1.fastq.gz> </PATH/TO/FOLD/INPUT _bulk2_2.fastq.gz>

3. Alineación a la secuencia de referencia

/PATH/TO/APP/bowtie2 -p 16 -x ./PATH/TO/FOLD/REFERENCE _GENOME--very-sensitive-local -1 /PATH/TO/FOLD/INPUT_bulk1_1_val_1.fq -2 /PATH/TO/FOLD/INPUT_bulk1 _2_val_2.fq -S /PATH/TO/FOLD/OUTPUT_bulk1.sam

/PATH/TO/APP/bowtie2 -p 16 -x ./PATH/TO/FOLD/ REFERENCE _GENOME --very-sensitive-local -1 /PATH/TO/FOLD/OUTPUT_bulk2_1_val_1.fq -2 /PATH/TO/FOLD/OUTPUT_bulk2 _2_val_2.fq -S PATH/TO/FOLD/OUTPUT_bulk2.sam

4. Ordenar archivos con picard y convertir sam a bam (SortSam- Picard tool)

java -Xmx64g -jar /PATH/TO/APP/picard.jar SortSam INPUT=/PATH/TO/FOLD/INPUT_bulk1.sam OUTPUT=/PATH/TO/FOLD/OUTPUT_bulk1_sorted.bam SORT_ORDER=coordinate\

java -Xmx64g -jar /PATH/TO/APP/picard.jar SortSam INPUT=/PATH/TO/FOLD/INPUT_bulk2.sam OUTPUT=/PATH/TO/FOLD/OUTPUT_bulk2_sorted.bam SORT_ORDER=coordinate\

5. Agregar etiquetas a todos en formato ordenado bam (AddOrReplaceReadGroups-Picard tool)

java -Xmx64g -jar /PATH/TO/APP/picard.jar AddOrReplaceReadGroups INPUT=/PATH/TO/FOLD/INPUT/INPUT_bulk1_sorted.bam OUTPUT=/PATH/TO/FOLD/OUTPUT/OUTPUT_bulk1_RG_sorted.bam RGID= ReadGroupID RGSM= ReadGroupSamplename RGLB=ReadGroupLibrary RGPL=ILLUMINA RGPU=ignore VALIDATION_STRINGENCY=LENIENT

java -Xmx64g -jar /PATH/TO/APP/picard.jar AddOrReplaceReadGroups INPUT=/PATH/TO/FOLD/INPUT/INPUT_bulk2_sorted.bam OUTPUT=/PATH/TO/FOLD/OUTPUT/OUTPUT_ bulk2_RG_sorted.bam RGID= ReadGroupID RGSM= ReadGroupSamplename RGLB=ReadGroupLibrary RGPL=ILLUMINA RGPU=ignore VALIDATION_STRINGENCY=LENIENT

6. Marcar duplicados en archivos RG bam ordenados (MarkDuplicates- Picard tool)

java -Xmx64g -jar /PATH/TO/APP/picard.jar MarkDuplicates INPUT=/PATH/TO/FOLD/INPUT/INPUT_bulk1_RG_sorted.bam OUTPUT=/PATH/TO/FOLD/OUTPUT/OUTPUT_bulk1_mkdupl.bam METRICS_FILE PATH/TO/FOLD/ OUTPUT_bulk1_mkdupMetrics.txt

java -Xmx64g -jar /PATH/TO/APP/picard.jar MarkDuplicates INPUT=/PATH/TO/FOLD/INPUT/INPUT_bulk2_RG_sorted.bam OUTPUT=/PATH/TO/FOLD/OUTPUT/OUTPUT_bulk2_mkdupl.bam METRICS_FILE PATH/TO/FOLD/ OUTPUT_bulk2_mkdupMetrics.txt

---

Es posible solicitar estadísticas y gráficos que describan el estado de los archivos .bam mediante el programa Qualimap.

/PATH/TO/APP/qualimap bamqc -bam /PATH/TO/FOLD/INPUT/INPUT_bulk1_mkdupl.bam --java-mem-size=4G -outdir /PATH/TO/FOLD/OUTPUT -outfile NAME.pdf -c -nw 400 -hm 3

/PATH/TO/APP/qualimap bamqc -bam /PATH/TO/FOLD/INPUT/INPUT_bulk2_mkdupl.bam --java-mem-size=4G -outdir /PATH/TO/FOLD/OUTPUT -outfile NAME.pdf -c -nw 400 -hm 3

Alineado de secuencias genómicas:

7. Indexar archivos de referencia y bam

a) Indexar archivos de referencia

export LD_LIBRARY_PATH=/share/apps/zlib/lib/:${LD_LIBRARY_PATH}

/PATH/TO/APP/samtools/samtools faidx /PATH/TO/FOLD/INPUT/INPUT/S_lycopersicum_chromosomes.3.00.fa

b) También se necesita crear un diccionario para el genoma de referencia

java -Xmx64g -jar /PATH/TO/APP/picard.jar CreateSequenceDictionary
R=/PATH/TO/FOLD/REFERENCE_GENOME.fa O=/PATH/TO/FOLD/
REFERENCE_GENOME.dict

c) Indexar archivos bam

java -Xmx64g -jar /PATH/TO/APP/picard.jar BuildBamIndex
I=/PATH/TO/FOLD/INPUT/INPUT_ bulk1_mkdupl.bam

java -Xmx64g -jar /PATH/TO/APP/picard.jar BuildBamIndex
I=/PATH/TO/FOLD/INPUT/INPUT_ bulk2_mkdupl.bam

8. Detección de variantes usando GATK:

a) Detección de variantes en cada grupo o *bulk*

/PATH/TO/APP/gatk --java-options "-Xmx64g" HaplotypeCaller –R
/PATH/TO/FOLD/REFERENCE_GENOME.fa -I
/PATH/TO/FOLD/INPUT/INPUT_bulk1_mkdupl.bam -ERC GVCF -O
/PATH/TO/FOLD/OUTPUT/OUTPUT_bulk1_raw_variants_gvcf.g.vcf

/PATH/TO/APP/gatk --java-options "-Xmx64g" HaplotypeCaller –R
/PATH/TO/FOLD/REFERENCE_GENOME.fa -I
/PATH/TO/FOLD/INPUT/INPUT_bulk2_mkdupl.bam -ERC GVCF -O
/PATH/TO/FOLD/OUTPUT/OUTPUT_bulk2_raw_variants_gvcf.g.vcf

b) Combinar los resultados en un único archivo VCF con los datos “crudos”:

I) Combinar dos o más archivos GVCFs (HaplotypeCaller-GATK)

/PATH/TO/APP/gatk CombineGVCFs -R /PATH/TO/FOLD/REFERENCE_GENOME.fa --
variant /PATH/TO/FOLD/ INPUT/INPUT_bulk1_raw_variants_gvcf.g.vcf--variant
/PATH/TO/FOLD/INPUT/INPUT_bulk2_raw_variants_gvcf.g.vcf -O
/PATH/TO/FOLD/OUTPUT/combinedbulks_cohort.g.vcf.gz

II) Crear un archivo VCF solo con las variantes (GenotypeGVCFs-GATK)

/PATH/TO/APP/gatk --java-options "-Xmx2g" GenotypeGVCFs -R
/PATH/TO/FOLD/REFERENCE_GENOME.fa -V /PATH/TO/FOLD/INPUT/
combinedbulks_cohort.g.vcf.gz -O /PATH/TO/FOLD/OUTPUT/
combined_raw_InDels_and_SNPs.vcf

c) Recalibrar las variantes con GATK

I. Extraer los *SNP*s del archivo con variantes (SelectVariants-GATK)

/PATH/TO/APP/gatk SelectVariants -R /PATH/TO/FOLD/REFERENCE_GENOME.fa -V /PATH/TO/FOLD/INPUT/combined_raw_InDels_and_SNP.vcf --select-type-to-include *SNP* -O /PATH/TO/FOLD/OUTPUT/combined_raw_SNPs.vcf

II. Extraer los *INDEL* del archivo con variantes (SelectVariants-GATK)

/PATH/TO/APP/gatk SelectVariants -R /PATH/TO/FOLD/REFERENCE_GENOME.fa -V /PATH/TO/FOLD/INPUT/combined_raw_InDels_and_SNP.vcf --select-type-to-include *INDEL* -O /PATH/TO/FOLD/OUTPUT/combined_raw_InDels.vcf

III. Determinar los parámetros de filtrado para los *SNP*s (VariantFiltration-GATK)

/PATH/TO/APP/gatk gatk VariantFiltration -R /PATH/TO/FOLD/REFERENCE_GENOME.fa -V /PATH/TO/FOLD/INPUT/combined_raw_SNPs.vcf -O /PATH/TO/FOLD/OUTPUT/combined_filtered_SNPs.vcf --filter-expression 'QD < 2.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MappingQualityRankSum < -12.5 || ReadPosRankSum < -8.0' --filter-name "my_*SNP*_filter"

IV. Determinar los parámetros de filtrado para los *InDel* (VariantFiltration-GATK)

/PATH/TO/APP/gatk gatk VariantFiltration -R /PATH/TO/FOLD/REFERENCE_GENOME.fa -V /PATH/TO/FOLD/INPUT/combined_raw_*SNP*s.vcf -O /PATH/TO/FOLD/OUTPUT/combined_filtered_*InDel*s.vcf --filter-expression 'QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0' --filter-name "my_*InDel*_filter"

9. Filtrar los *SNP*s e *INDEL* (Filtrar las muestras en un solo archivo vcf para obtener un archivo "pass" para los *SNP*s o *InDel*s)

/PATH/TO/APP/vcftools --vcf PATH/TO/FOLD/INPUT/combined_filtered_*SNP*s.vcf --remove-filtered-all --recode --recode-INFO-all -c > PATH/TO/FOLD/OUTPUT/combined_filtered_*SNP*s_only_pass.vcf

/PATH/TO/APP/vcftools --vcf PATH/TO/FOLD/INPUT/combined_filtered_*InDel*s.vcf --remove-filtered-all --recode --recode-INFO-all -c > PATH/TO/FOLD/OUTPUT/combined_filtered_*InDel*s_only_pass.vcf

Identificación de regiones genómicas asociadas al carácter tipo de carpelo:

Descargar e instalar el paquete de R devtools

install.packages("devtools")

library ("devtools")

Descargar el paquete QTLseqr con devtools

```
devtools::install_github("bmansfeld/QTLseqr")
#cargar el paquete
library ("QTLseqr")
# cargar el paquete
library (readr)
# Configurar la muestra y el nombre del archivo
rawdata <- "C: PATH/TO/FOLD"
HighBulk <- "unfused_bulk_ID"
LowBulk <- "fused_bulk_ID"
Chroms <- paste0(rep("SL3.0ch", 12),
c("01","02","03","04","05","06","07","08","09","10","11","12") )
#Importar los datos de SNP desde el archivo
SNP_data <-
  importFromGATK(
    file = rawdata,
    highBulk = HighBulk,
    lowBulk = LowBulk,
    chromList = Chroms
)
#Filtar los SNPs
library("ggplot2")
setwd ("C: PATH/TO/FOLD")
### Analizar la profundidad de lectura de los datos
ggplot(data = SNP_data) +
  geom_histogram(aes(x = DP.HIGH + DP.LOW)) +
  xlim(0,250)
ggsave( filename = "DPTotal.pdf, dpi = 300)
### Analizar la frecuencia del alelo de referencia
ggplot(data = SNP_data) +
  geom_histogram(aes(x = REF_FRQ))
ggsave( filename = "ReferenceFrequency.pdf", dpi = 300)
```

```
### Analizar el índice-SNP index en cada bulk
ggplot(data = SNP_data) +
  geom_histogram(aes(x = SNPindex.HIGH)) #unfused bulk
ggsave( filename = "SNPindex.HIGH_SL3.0.pdf", dpi = 300)
ggplot(data = SNP_data) +
  geom_histogram(aes(x = SNPindex.LOW)) #fused bulk
ggsave( filename = "SNPindex.LOW_SL3.0.pdf", dpi = 300)
#Filtrar los SNPs en base a algún criterio
df <-
  filterSNPs(
    SNPset = SNP_data,
    refAlleleFreq = 0.20,
    minTotalDepth = 38,
    maxTotalDepth = 65,
    depthDifference = 50,
    minSampleDepth = 5,
    minGQ = 99,
    verbose = TRUE
)
# Exportar los datos de SNPs a una tabla
write.csv (SNP_data, row.names = F, file= "NAME.csv")
write.csv (df, row.names = F, file= "NAME.csv")
#Analisis según Takagi et al.(2013)_qtl_seq
df <- runQTLseqAnalysis(df,
                        windowSize = 2e6,
                        popStruc = "F2",
                        bulkSize = 10,
                        replications = 10000,
                        intervals = c(95, 99)
)
#Análisis según Magwene et al. (2011)_G
```

```
df <- runGprimeAnalysis(df,
                        windowSize = 2e6,
                        outlierFilter = "deltaSNP",
                        filterThreshold = 0.1)
```

Esquematización de *QTL* putativos y exportación de datos:

```
# Gráfico de distribución de los valores de G'
plotGprimeDist(SNPset = df, outlierFilter = "Hampel")
plotGprimeDist(SNPset = df, outlierFilter = "deltaSNP", filterThreshold = 0.1)

# Gráfico del análisis de QTL_seq
p1 <- plotQTLStats(SNPset = df_filt, var = "nSNPs")
p1
p2 <- plotQTLStats(SNPset = df_filt, var = "deltaSNP", plotIntervals = TRUE)
p2
p3 <- plotQTLStats(SNPset = df_filt, var = "Gprime", plotThreshold = TRUE, q = 0.01)
p3
QTLplots <- plotQTLStats(
  SNPset = df,
  var = "negLog10Pval",
  plotThreshold = TRUE,
  q = 0.01,
  subset = c("SL3.0ch03""SL3.0ch06", "SL3.0ch10")
)
QTLplots
#Exportar los datos de QTLs putativos en formato csv
results<- getQTLTable(SNPset = df, method = "Gprime", alpha = 0.01, export = TRUE,
fileName = "QTL_Gprime.csv")
results
```

Tabla S1: Fragmento de archivo con extensión .table devuelto por el programa GATK.SL95824: bulk no fusionado. SL95825: bulk fusionado. CHROM: cromosoma, POS: posición física del polimorfismo, REF: alelo de referencia, ALT: alelo alternativo, AD: cantidad de lecturas para cada alelo, DP: profundidad de cobertura total, QC: calidad del genotipo, PL: probabilidades con escala de Phred

CHROM	POS	REF	ALT	SL95824.AD	SL95824.DP	SL95824.GQ	SL95824.PL	SL95825.AD	SL95825.DP	SL95825.GQ	SL95825.PL
SL3.0ch00	4314	A	C	2,29	31	73	1101,73,0	0,33	33	98	1258,98,0
SL3.0ch00	38404	G	A	6,17	23	99	617,0,171	19,13	32	99	436,0,671
SL3.0ch00	70564	T	A	0,37	37	99	1487,111,0	0,27	27	81	1061,81,0
SL3.0ch00	98676	A	T	4,2	6	55	55,0,162	12	12	33	0,33,495
SL3.0ch00	98680	C	T	4,2	6	55	55,0,162	12	12	9	0,9,336
SL3.0ch00	98683	C	T	4,2	6	55	55,0,162	12	12	33	0,33,495
SL3.0ch00	100944	T	C	0,38	38	99	1508,114,0	0,22	22	66	890,66,0
SL3.0ch00	107045	G	C	24,12	36	99	393,0,814	13,2	33	99	720,0,440
SL3.0ch00	133828	T	A	20,9	29	99	279,0,723	12,14	26	99	494,0,403
SL3.0ch00	135824	G	A	0,36	36	99	1486,108,0	0,22	22	66	905,66,0
SL3.0ch00	136252	G	A	16,25	41	99	745,0,479	10,1	20	99	338,0,326
SL3.0ch00	179893	T	C	32,6	38	99	156,0,1267	13	13	39	0,39,532
SL3.0ch00	179910	T	A	34,7	41	99	186,0,1462	10	10	30	0,30,398
SL3.0ch00	188721	T	G	0,21	21	62	776,62,0	0,15	15	45	580,45,0
SL3.0ch00	196772	A	G	0,31	31	91	1138,91,0	0,26	26	78	1043,78,0
SL3.0ch00	228550	G	C	0,26	26	78	996,78,0	0,25	25	75	996,75,0