UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE CIENCIAS

# IDENTIFICACIÓN DE SNPs MEDIANTE GENOTIPADO POR SECUENCIACIÓN PARA EL MEJORAMIENTO GENÉTICO DE TRIGO (*Triticum aestivum* L.)

**por**

**Bettina LADO LINDNER**

Orientador: PhD. Jarislav VON ZITZEWITZ

Co-Orientador: PhD. François BELZILE

**TESIS presentada como uno de los requisitos para obtener el título de *Magister* en Biotecnología**

Montevideo, Uruguay

Diciembre 2012

Tesis aprobada por el tribunal integrado por el Ing. Agr. (PhD) Ignacio Aguilar, Ing. Agr. (PhD) Lucia Gutiérrez, Ing. Agr. (PhD) Omar Borsani y el Lic. (PhD) Sabina Vidal el día 5 de Diciembre del 2012. Autor: Bettina Lado. Director: BSc. (PhD) Jarislav von Zitzewitz. Co-Director: BSc. (PhD) Fraçois Belzile.

**Agradecimientos**

A mi tutor, Jarislav von Zitzewitz, por haberme guiado en este trabajo y estar cuando era necesario, porque fue un camino de aprendizaje para ambos yo como estudiante y él como tutor, por que supo ser un muy buen tutor y sobre todo buena persona.

A mi co-tutor François Belzile, por haberme re-direccionado en mi trabajo y ser la guía que necesitaba en un momento de este largo camino. Por haberme abierto las puertas y hecho sentir tan a gusto durante los meses que pase en su laboratorio en la Universidad Laval, Quebec. Gracias a todo su equipo, quienes formaron parte de una experiencia muy importante en este recorrido, Aurelie, Humira, Elmer, Maxime, Gabrielle y Martin.

A Omar Borsani y el Laboratorio de Bioquímica de Facultad de Agronomía por haberme abierto las puertas y apoyado principalmente en los comienzos de este recorrido, donde más lo necesite. Por su buena onda y el tiempo compartido.

A Ivan Matus y Luis Inostroza quienes aportaron los datos fenotípicos medidos a campo en Chile.

A Joanna Lado y Matías Manzi por su contribución y tutoría informal en este trabajo, y sobre toda las cosas por el "aguante de cabeza" en algunos de los momentos más críticos de este proceso.

A los compañeros de INIA "La Estanzuela" por hacer muchos momentos de ardua labor fáciles de llevar. Gracias también al apoyo por parte de algunos investigadores de INIA, principalmente a Federico Condón y Martín Quincke que aportaron su cuota de conocimiento ha este trabajo.

A mi compañeras de Laboratorio Paula Silva y Luciana Landechea por colaborar de diversas formas siempre que fue necesario, gracias.

A los amigos de la vida, que me recargan las energías y renuevan las motivaciones que me permiten seguir adelante en los momentos difíciles.

A mi familia por su apoyo incondicional, una vez más. Por ser la fuerza de movimiento en lo momentos en que era más fácil bajar los brazos y por mostrarme que el sacrificio muchas veces es necesario para alcanzar nuestras metas. Porque son la piedra fundamental para que esto haya sido posible.

Finalmente este trabajo no podría haber sido logrado sin el financiamiento del FONDO REGIONAL DE TECNOLOGÍA AGROPECUARIA (FONTAGRO)-8038, titulado: “Aumento de la competitividad de los sistemas productivos de papa y trigo en Sudamérica ante el cambio climático”, ni sin el apoyo del Instituto Nacional de Investigación Agropecuaria (INIA), gracias.

TABLA DE CONTENIDO

Página

PÁGINA DE APROBACIÓN……...…….....II

AGRADECIMIENTOS.....III

RESUMEN.....VII

SUMMARY.....VIII

ABREVIATURAS.....IX

1. INTRODUCCION.....10
   1.1. BREVE HISTORIA DE LA AGRICULTURA Y EL MEJORAMIENTO GENÉTICO.....10
   1.2. CULTIVO DE TRIGO (*Triticum aestivum* L.).....11
   1.3. MÉTODOS DE MEJORA GENÉTICA.....12
      1.3.1. MARCADORES MOLECULARES.....13
         1.3.1.1. Identificación de SNPs .....14
   1.4. SNPs EN EL ANÁLISIS DE ESTRUCTURA DE POBLACIONES Y ANALISIS DE DESEQUILIBRIO DE LIGAMIENTO.....21
   1.5. ASOCIACIONES ENTRE GENOTIPO Y FENOTIPO.....22
   1.6. SELECCIÓN GENOMICA Y MEJORAMIENTO VEGETAL.....25
   1.7. OBJETIVOS.....27
2. METODOLOGÍA.....28
   2.1. GENOTIPADO.....28
   2.2. IDENTIFICACIÓN DE MARCADORES.....28
      2.2.1. Pipeline de Tassel.....29
      2.2.2. Pipeline con genoma de referencia.....30
   2.3. EVALUACIÓN DE LAS METODOLOGÍAS UTILIZADAS EN LA IDENTIFICACIÓN DE SNPs (PIPELINES).....31
   2.4. VALIDACIÓN DE SNPs.....31
   2.5. Comparación entre matrices genéticas con sub-grupo de datos.....31

2.6. Análisis de estructura de la población.....32
2.7. Comparación entre matrices de coancestría y de parentesco con marcadores.....32
2.7.1. Mapeo de SNPs.....32
Análisis de desequilibrio de ligamiento.....33
2.7.2. Predicciones fenotípicas.....33
3. RESULTADOS.....36
3.1. GENOTIPADO.....36
3.2. COMPARACIÓN ENTRE PIPELINES.....38
3.3. VALIDACIÓN DE LOS SNPs IDENTIFICADOS.....40
3.3.1. Análisis de correlación entre matrices genéticas.....41
3.3.2. Análisis de estructura de la población.....42
3.3.3. Mapeo de SNPs y análisis de desequilibrio de ligamiento.....43
3.3.4. Predicciones fenotípicas.....46
4. DISCUSIÓN .....50
4.1. IDENTIFICACIÓN DE SNPs.....51
4.2. VALIDACIÓN DE SNPs.....53
4.2.1. Correlación entre matrices genéticas.....54
4.2.2. Correlación entre marcadores y pedigrí.....54
4.2.3. Mapeo de SNPs.....55
4.2.4. Predicciones fenotípicas.....56
5. CONCLUSIONES.....59
6. BIBLIOGRAFÍA.....60
7. ANEXOS.....78

## RESUMEN

En mejoramiento genético de cultivos la posibilidad de predecir comportamientos fenotípicos de los distintos cultivares en el campo se ha incrementado debido a los recientes avances en las tecnologías de análisis molecular. Sin embargo, en el caso de trigo (*Triticum aestivum*), existen limitaciones para aplicar algunas de las nuevas tecnologías de identificación de marcadores moleculares, debido la complejidad de su genoma. En este sentido, el objetivo de este trabajo es aplicar y validar una nueva metodología para identificar variaciones en una sola base (SNPs) en el genoma de trigo. Dicha metodología es sencilla y comprende la construcción y secuenciación de librerías genómicas, a través de la técnica conocida como "Genotipado por Secuenciación", permitiendo reducir costos en relación a otros métodos. Las muestras de ADN genómico son secuenciadas para posteriormente realizar la identificación de SNPs a través del análisis bioinformático de las secuencias, por dos metodologías diferentes. En la primera (PipT) se comparan entre sí los fragmentos de ADN de las distintas muestras para identificar posibles variaciones entre ellas y en la segunda (PipRef), la comparación se realiza con un ensamblado del genoma de trigo disponible. El método PipT fue el método seleccionado debido a que identificó SNPs con mejores características. El filtrado de falsos SNPs es un paso fundamental en el proceso. Se identificaron 28.199 SNPs en 382 muestras de trigo. Estos SNPs se utilizaron para realizar predicciones fenotípicas de un carácter con baja heredabilidad (rendimiento) y un carácter con alta heredabilidad (peso de mil granos), en dos condiciones ambientales contrastantes: riego y secano. Esto permitió comparar los resultados obtenidos con los datos reales, previamente conocidos, por el método de validaciones cruzadas. Los datos fenotípicos presentaron variación espacial, por lo que se evaluaron distintos modelos para corregir la tendencia encontrada, seleccionando un modelo de ajuste que utiliza medias móviles como covariable. La correlación de las predicciones en rendimiento ajustando por bloques incompletos (diseño a campo) fue de 0,354 y 0,311, obteniéndose una mejora significativa al aplicar el ajuste espacial, en el cual los valores ascendieron a 0,396 y 0,357 en riego y secano, respectivamente. Para peso de mil granos, carácter con mayor heredabilidad, las correlaciones fueron de 0,751 y 0,748 con el ajuste del diseño, frente a 0,787 y 0,799 con ajuste espacial, en riego y secano, respectivamente. Estos resultados permiten concluir que el método desarrollado puede ser aplicado en forma efectiva para detección de SNPs en trigo y que los mismos son válidos tanto para la predicción del peso de mil granos como del rendimiento. Lo anterior constituye una herramienta muy

potente que podría ser utilizada en la selección de material genético dentro de los programas de mejoramiento genético de trigo.


## SUMMARY

In crop breeding, the interest of predicting phenotypic behaviors of cultivars in the field has increased due to recent advances in molecular breeding technologies. However the complexity of wheat genome presents some limitation for applied new technologies in molecular markers identification with Next Generation Sequencing (NGS) technologies. Therefore, we applied "Genotyping by Sequencing" (GBS), a recently developed method to identify single nucleotide polymorphisms (SNPs), in the genome of 382 individuals of wheat (*Triticum aestivum*). GBS involves the construction and sequencing of genomic libraries; a simple and low-cost method directly applicable to plant and animal breeding programs. Genomic DNA samples were sequenced to subsequently perform the identification of SNPs through two different bioinformatic analysis methods, which compare the DNA fragments of different samples to identify variations between them or against a pseudo reference genome of wheat. The first one identified SNPs with better characteristics and therefore was used for all subsequent analysis. The step in filtering false SNPs was essential in deciding between these two analyses. 28,199 SNPs were identified in 382 samples of wheat. These SNPs were used to predict a low heritable phenotype (yield) and a highly heritable phenotype (thousand kernel weight), derived from two fields with two different environmental conditions: irrigated and without irrigation. This allowed comparing genomic predictions with the actual data by performing cross-validations. The phenotypic data showed spatial variation. Therefore, different models were tested to correct the trend found in the field, where the model that best fits the data was the one using moving-means as a covariate. The correlations of the predicted yields, adjusted for the incomplete block field design, were 0.354 and 0.311 and with spatial adjustment were 0.396 and 0.357, in irrigated and without irrigation, respectively. For thousand kernel weight, correlations were 0.751 and 0.748 with field design adjustment and 0.787 and 0.799 with spatial adjustment in irrigated and without irrigation respectively. These results allow concluding that the developed method can be applied effectively for SNPs detection in wheat being valid for thousand kernel weight and yield predictions. Finally, this is a powerful tool that could be applied in selection of genetic materials inside a breeding program.

**ABREVIATURAS**

SNP – Polimorfismo en una sola base de ADN

NGS – Secuenciación por técnicas de nueva generación

GBS – Genotipado por secuenciación

PCA – Análisis de componentes principales

PCoA – Análisis de coordenadas principales

LD – Desequilibrio de ligamiento

GS – Selección Genómica

GEBV – Valores de cría genéticos estimados

PipT – Pipeline de Tassel

PipRef – Pipeline con genoma de referencia

LS – Diseño de Látice Simple

FC – Diseño de Filas y Columna

BCA_MVNG – Diseño de bloques completos al azar con medias móviles como covariable

MVNG – Diseño de regresión lineal con medias móviles como covariable

# 1. INTRODUCCIÓN

## 1.1. BREVE HISTORIA DE LA AGRICULTURA Y EL MEJORAMIENTO GENÉTICO

La agricultura puede definirse como "arte, ciencia e industria que se ocupa de la explotación de plantas y animales para el uso humano". La agricultura moderna depende en gran medida de otras ciencias como la ingeniería, ciencias biológicas y físicas, así como también de la tecnología disponible. Existen diversas teorías de las causas que impulsaron el desarrollo de la agricultura. Una de ellas habla de causas multivariadas en las cuales intervienen la presión por la necesidad de recursos (presiones poblacionales), los cambios climáticos y el desarrollo tecnológico (Flannery, 1982; McNeish, 1992). Se estima que la agricultura como tal se inicia en el año 10.000 a.C., y luego de 3000 años se convierte en la principal forma de obtener alimentos en diversas sociedades en todo el mundo. La agricultura no sólo permitió el crecimiento de las poblaciones, dada la mayor abundancia de alimentos, sino que permitió la conquista de nuevos territorios en los cuales no se podía vivir anteriormente.

Los primeros cultivos se domesticaron en Medio Oriente entre los años 9600 y 8500 a.C., siendo el trigo (*Triticum monoccocum y T. diccocum*) y la cebada (*Hordeum vulagre*) los primeros cereales en ser domesticados. En China, en el año 7500 a.C., aparecen los primeros indicios de cultivo de arroz (*Oryza sativa*) y soja (*Glycine max*) (Diamond, 2002). La domesticación de plantas implica modificar artificialmente un esquema genético generado a través de la selección natural, con el objetivo de adaptar las plantas a las distintas características de las zonas habitadas por el hombre. Esta selección favorece la sobrevivencia de características deseadas en estas poblaciones de cultivos, en detrimento de las indeseables. Por lo tanto, la domesticación es definida como un proceso de selección genética continuo llevado a cabo por el hombre. (Gepts y Papa, 2002; Gepts, 2004).

El mejoramiento genético vegetal puede definirse como la ciencia cuyo objetivo es modificar genotipos, mejorándolos con un fin determinado de acuerdo a las necesidades del hombre (Frankel, 1958). El principal objetivo de la domesticación de cultivos ha sido el incremento del rendimiento, con el fin de aumentar la disponibilidad de alimento para el ser humano. Existen ciertas condiciones necesarias dentro de un proceso de mejora genética vegetal, entre las que destacan, la existencia de variabilidad genética o la capacidad de crearla, la capacidad para detectar dichas diferencias con potencial valor para el hombre y la manipulación de

dicha variación que permita obtener cultivares agronómicamente estables (Zamir, 2001). Los mayores avances en el mejoramiento genético de plantas se han producido en la era moderna, debido a algunos descubrimientos científicos claves. Las primeras evidencias son aportadas en 1694 por Camerarius, quien describe la sexualidad en las plantas (Zirkle, 1935 revisado por Rieseberg y Carney, 1998). Luego en 1717, Fairchild produce artificialmente la primera planta híbrida de clavel (Rieseberg y Carney, 1998) y en 1859 Darwin publica el "Origen de la especies", libro en el cual postula su teoría sobre la selección natural y la evolución. Posteriormente, en 1866, Mendel realiza sus primeros experimentos que dieron origen a las conocidas "Leyes de Mendel". A partir de estos conocimientos, ya en el siglo XX se producen un gran número de descubrimientos en el área de la genética y el mejoramiento vegetal, los cuales han aportado las bases para entender los mecanismos que operan en la selección natural (Bringham, 1981; Borlaug, 1983). Dichos descubrimientos han aportado información para determinar los patrones de evolución de las especies cultivadas, variaciones mendelianas, hibridación intra-específica y poliploidía, estos patrones que se han utilizado por los mejoradores son los mismos mecanismos que han formado parte de la selección natural (Vaughan et al., 2007)

Una de las mayores contribuciones de los procesos de mejora genética ha sido la adaptación de variedades a nuevas áreas, lo que requiere ajustar el ciclo de crecimiento y los requerimientos del cultivo a las características de la nueva región. Otros aportes importantes han sido la mejora en características agronómicas, entre las que se destacan resistencia a enfermedades y plagas, la tolerancia a condiciones ambientales adversas como por ejemplo, altas y bajas temperaturas y déficit hídrico (Cornide, 2001).

## 1.2. CULTIVO DE TRIGO (*Triticum aestivum* L.)

El trigo es uno de los principales cultivos a nivel mundial, siendo el pilar básico para la alimentación humana. Actualmente la producción mundial de trigo se ubica en el quinto lugar con 653 millones de tonelada y ocupa el primer lugar en hectáreas sembradas (FAO-STAT, 2010). Los principales países productores son: China, India, Estados Unidos, Francia y la Federación Rusa, en el hemisferio sur Australia, Argentina y Brazil (FAO-STAT, 2010). En Uruguay, el trigo es el principal cultivo de invierno, y su producción ha ido en aumento en los últimos años, con una producción de 205 mil toneladas en la zafra 2002/03 a 1.300 miles

de toneladas en 2010/11 (DIEA-MGAP, 2011). Esto lo convierte en el segundo cultivo en importancia después de la soja (FAO-STAT, 2010).

El trigo es una especie autógama y hexaploide, producto de la hibridación de 2 especies de *T. turgidum* (el cual posee los genomas A y B) y *Aegilops tauschii* (que aporta el genoma D) (McFadden y Sears 1946, Kihara et al., 1944 revisado por Chao et al., 2010). A su vez, la especie *T. turgidum* es producto de una hibridación previa de *T. uratu* (genoma A) y *A. speltoides (*genoma B) (Sarkar y Stebbins, 1956; Dvorák y Zhang, 1990; Dvorák et al., 1993). En este sentido, la característica poliploide del trigo ha dificultado el desarrollo de mapas genéticos, así como el estudio de caracteres de herencia compleja, ya que las variables agronómicas se ven influenciadas por genes provenientes de 3 genomas distintos. Algunas de estas dificultades han sido superadas con la construcción de plantas nulisómicas (Lafever y Patterson, 1964) y monosómicas (Allan y Vogel, 1964), las cuales poseen un menor número de cromosomas que las plantas normales. A su vez, el genoma de trigo presenta otras características, como su gran tamaño (16Gb, cinco veces el genoma humano), un alto número de regiones repetidas (80%) (Bennett y Smith, 1976;) y un 25 a 30 % de sus genes duplicados (Dubcovsky et al., 1996; Akhunov et al., 2003) que dificultan el desarrollo de análisis genéticos precisos en esta especie. Posee una gran ventaja ya que, al ser una especie autógama, presenta bajos niveles de heterocigocidad y bajas proporciones de flujo génico (<1 %) debido a un reducido nivel de polinización cruzada (Martin, 1990; Gustafson et al., 2005).

Uruguay fue el país pionero en América Latina en lo referente al mejoramiento del cultivo de trigo con los trabajos desarrollados por el Dr. Alberto Boerger (Bonfanti, 2010). La creación del Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) permitió intensificar la colaboración entre México y otros países y el envío de materiales al resto del mundo (Heisey et al., 2002). Por lo tanto, la investigación en mejoramiento genético en el cultivo de trigo en Uruguay es un pilar fundamental que aporta al desarrollo económico del país y del mundo.

### 1.3. MÉTODOS DE MEJORA GENÉTICA

Los métodos clásicos de mejora genética de cultivos (cruzamiento, retrocruzamiento y selección a campo de genotipos más promisorios) han sido y siguen siendo fundamentales en la obtención de cultivares con las características fenotípicas y agronómicas deseadas

(Koebner y Summers, 2003). Por otro lado se han desarrollado nuevos métodos que apuntan a la creación de variabilidad genética. En 1920 se descubre la posibilidad de generar mutaciones en el genoma, exponiendo la planta a radiaciones o por medio de compuestos químicos, lo que permitió obtener variedades comerciales empleadas actualmente (Peterson y Shama, 2005). Este momento podría considerarse el comienzo de la incorporación de técnicas biotecnológicas o de ingeniería genética en la mejora clásica de los cultivos. Estas herramientas ofrecen diversas ventajas, entre las que se destacan la inclusión de genes de interés presentes en otras especies dentro del cultivo objetivo del mejoramiento (transgénesis) y la posibilidad de acortar los tiempos en la obtención de un cultivar con las características deseadas. La selección asistida por marcadores moleculares se incluye dentro del grupo de métodos que permiten reducir estos tiempos, y se basan en la identificación de marcadores asociados a genes de interés, lo que permite seleccionar materiales por sus características genéticas (Koebner y Summers, 2003). El éxito de la selección mediante características genéticas depende en gran medida del número de genes involucrados en el control genético del carácter, de las relaciones inter-alélicas (dominancia o recesividad) y de la influencia del ambiente, medida comúnmente mediante el parámetro conocido como heredabilidad (Mackay et al., 2009). La heredabilidad se define como la proporción de variación fenotípica en una población que puede ser explicada por la varianza genética entre los individuos (Holland et al., 2003). El interés de la aplicación de estos marcadores dentro de un programa de mejoramiento genético, surge con diversos objetivos entre los que se destacan la posibilidad de identificar factores genéticos útiles en poblaciones, introgresar genes de interés en líneas de mejoramiento, estudiar la interacción genotipo ambiente, monitorear la diversidad genética e identificar cultivares y/o germoplasma con un mayor potencial agronómico (Sorrels, 1998; Stuber et al, 1999).

### 1.3.1. Marcadores Moleculares

Los marcadores moleculares surgen como una herramienta útil para alcanzar los objetivos de la mejora genética, permitiendo conjugar la variabilidad genotípica y fenotípica, siendo utilizados como criterio de selección (Bonnett et al., 2005). Hasta hace pocos años la única herramienta disponible para genotipar el ADN eran los marcadores del tipo microsatélites (SSR), polimorfismo en el largo de los fragmentos amplificados (AFLPs), polimorfismo en el largo de los fragmentos de restricción (RFLPs), entre otros. Estos métodos, requieren la

inversión de mucho tiempo y recursos, tanto para su desarrollo como para su aplicación en la evaluación de poblaciones, principalmente cuando es necesario evaluar un gran número de marcadores en diversos individuos (Gupta et al., 2008). Estos marcadores mostraron ser muy efectivos a la hora de identificar los efectos de genes que codifican características simples. En cambio, no sucedió así en el caso de características complejas, en las cuales interactúan muchos genes, cada uno de ellos en distinta proporción (Collard et al., 2005; Buerstymayr et al., 2009). Otro tipo de marcador empleado: los SNP, polimorfismos en una sola base. Este tipo de marcador ha cobrado importancia en los últimos tiempos debido a su abundancia en el genoma y a las tecnologías desarrolladas para su identificación (Gupta et al., 2008).

Dentro de las nuevas tecnologías que permiten identificar SNPs se encuentran los microarreglos (Jaccoud et al., 2001), que han significado un avance en la caracterización genotípica, dado que permiten una cobertura de alta densidad de marcadores y aceleran los tiempos de evaluación. Sin embargo, esta sigue siendo una tecnología muy costosa que requiere inversión de tiempo y recursos para ser desarrollada (Akbari et al., 2006). Estas limitantes han conducido al desarrollo de nuevas tecnologías que permiten el genotipado de miles de marcadores en cientos de individuos, sin la necesidad de una identificación previa de los mismos (Hyten et al., 2010c). Dentro de estas nuevas tecnologías se encuentran los métodos de preparación de librerías, la secuenciación y el análisis bioinformático para identificar SNPs, los cuales se basan en la re-secuenciación del genoma en pequeños fragmentos de ADN de las distintas muestras a genotipar. La alineación de estos fragmentos entre sí y/o contra un genoma de referencia, permite identificar variaciones genéticas de cambios en una única base nucleotídica, inserciones y/o deleciones de las mismas. Este tipo de secuenciación a gran escala permite identificar un elevado número de marcadores en el genoma, a un costo más reducido, haciendo posible su utilización como herramienta de apoyo a los programas de mejoramiento de cultivos vegetales (Ingvarsson y Street, 2011)

#### 1.3.1.1. Identificación de SNPs

**Métodos de preparación de librerías genómicas**

Existen diversas técnicas de preparación de librerías, las cuales, permiten preparar el ADN que va a ser secuenciado. Actualmente, los métodos más utilizados se centran en el uso de enzimas de restricción, las cuales realizan cortes en regiones específicas del genoma, que

permitan obtener una representación reducida de este. Existen distintos métodos que utilizan enzimas de restricción, los principales son, la secuenciación de representación reducida (RRL), la secuenciación de ADN asociada a los sitios de restricción (RAD-seq) y la secuenciación de baja cobertura para genotipado (Davey et al., 2011, Figura 1). Una de las grandes ventajas del uso de enzimas de restricción, es la posibilidad de excluir regiones altamente repetidas en el genoma, debido a que existen enzimas que son sensibles a las regiones metiladas y no son capaces de cortar en estos sitios (Davey et al., 2011). Todos los métodos de preparación de librerías con enzimas de restricción involucran la digestión del ADN genómico (con una o más de estas enzimas) y selección o reducción en el número de fragmentos, para la posterior secuenciación de los fragmentos seleccionados (Davey et al., 2011).

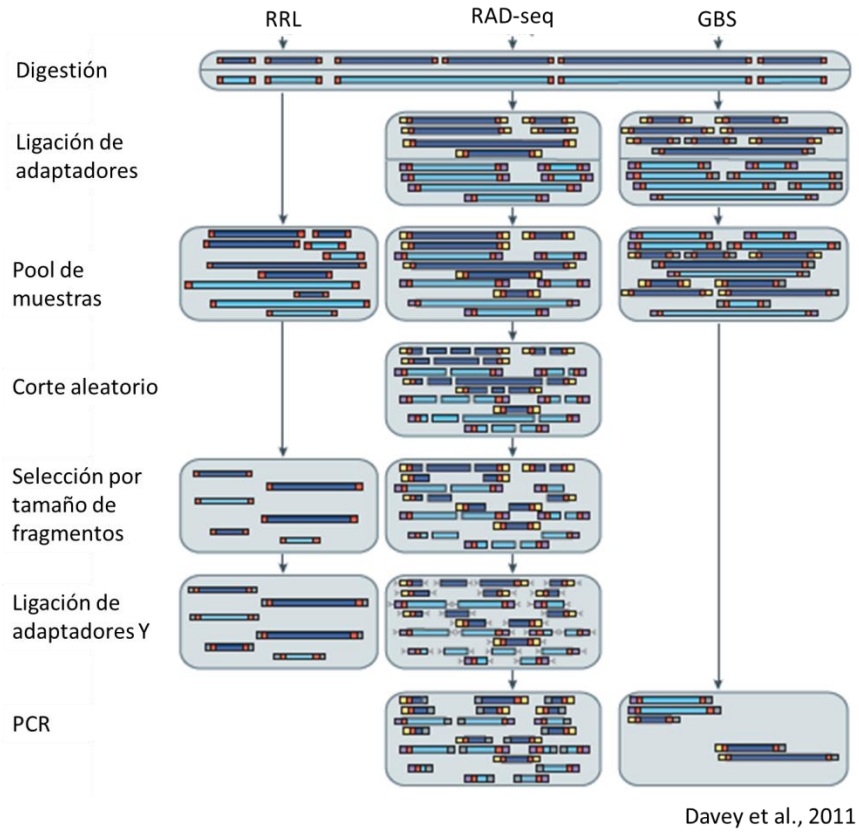



Figura1. Esquema de la preparación de librerías de ADN por distintos métodos que utilizan enzimas de restricción, secuenciación de representación reducida (RRL), secuenciación de ADN asociada a sitios de restricción (RAD-seq) y Genotipado por Secuenciación (GBS).

Este trabajo, se ha centrado en la evaluación de un método denominado Genotipado por Secuenciación (GBS), el cual se clasifica dentro del grupo de los métodos de secuenciación

de baja cobertura (Davey et al., 2011). Estos son métodos que permiten secuenciar un gran número de individuos, aunque con una menor cobertura en cada uno de ellos. Si bien en estos casos, se reduce la cobertura de marcadores por individuo, el método supone que un sub-grupo de diferentes marcadores son genotipados en distintos individuos. La estrategia entonces es poder utilizar marcadores que estén respaldados por un gran número de lecturas y mejorar la cobertura por medio de la imputación de los genotipos sin información dentro de la población de individuos que han sido secuenciados conjuntamente. Una ventaja muy importante de este método es su sencillez y el bajo costo por muestra (Elshire et al., 2011).

El primer paso en la aplicación de esta técnica es la obtención y posterior normalización de ADN de alta pureza. Este paso asegura que todas las muestras contienen la misma cantidad de ADN para ser secuenciado. A continuación, se colocan 2 adaptadores, los cuales van a ligarse al ADN. Estos adaptadores son oligonucleótidos que cuentan con: la secuencia complementaria al oligo que se encuentra unido a la celda de flujo del secuenciador, la secuencia de los oligonucleótidos usados para amplificar las lecturas y los sitios de restricción correspondientes. Uno de ellos cuenta también con el código de barras que permite identificar las muestras. Es importante que los adaptadores se encuentren a la misma concentración antes de ser adicionados en las distintas muestras. El siguiente paso consiste en la digestión del ADN con las enzimas de restricción, a continuación se agrega el adaptador junto con una enzima ligasa, ocurriendo la unión del ADN a los correspondientes adaptadores. Debido a que cada muestra se encuentra identificada con su código de barras, es posible secuenciar conjuntamente 96 o más muestras y formar el multiplex. A continuación, el pool de ADN se amplifica mediante la reacción en cadena de la polimerasa o Polymerase Chain Reaction (PCR) y se realiza una limpieza para eliminar posibles "artefactos". Para finalizar, la calidad de la librería se confirma por electroforesis en gel de agarosa y su concentración se mide con la técnica de PicoGreen™ (Molecular Probes, Eugene, OR) (Ahn et al., 1996) (Figura 2). Una vez alcanzado este punto, el material se secuencia masivamente. Esta tecnología de caracterización genotípica se ha aplicado recientemente con éxito en especies alógamas como maíz (Elshire et al., 2011), switchgrass (Grabowski et al., 2012), vid (Baldo et al., 2012) y eucaliptus (Grattapaglia, 2012), y también especies autógamas como ser cebada (Elshire et al., 2011) y trigo (Poland, et al. 2012a).

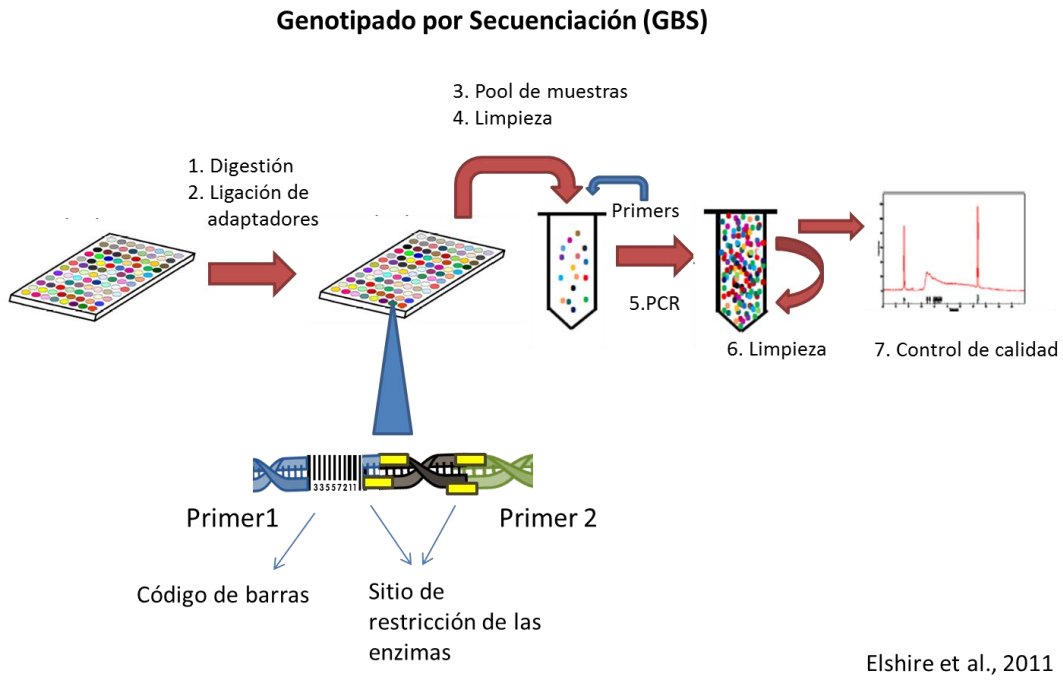



Figura 2. Esquema de la metodología de Genotipado por Secuenciación, en el cual se indica detalladamente los sucesivos pasos de la técnica.

**Métodos de secuenciación masiva**

La primera tecnología desarrollada con el fin de secuenciar el genoma ha sido el método de secuenciación automática Sanger (Sanger et al., 1977), la cual se aplicó para lograr la secuenciación del genoma humano (Lander et al., 2001; Venter et al., 2001). La tecnología Sanger se clasifica dentro del grupo de tecnologías de primera generación. La concentración de recursos en esta área permitió el desarrollo de nuevas tecnologías denominadas de secuenciación masiva o "Next Generation Sequencing" (NGS). Dentro de las NGS se pueden clasificar las tecnologías de segunda, tercera y cuarta generación. La principal ventaja de las tecnologías de segunda generación es su capacidad para generar mucha información con menos recursos, existiendo distintas plataformas desarrolladas para la secuenciación (Metzker et al., 2010; Niedringhaus et al., 2011). Las tecnologías de tercera y cuarta generación se diferencian de las de segunda generación porque se secuencia una única molécula de ADN (Levene et al., 2003). Lo anterior permite evitar el paso de amplificación, en el cual es posible generar errores de secuenciado. En el caso de las llamadas metodologías de cuarta

generación, existe también un cambio en el detector, en el cual se remplaza la fluorescencia por una señal eléctrica (Branton et al., 2008).

Este trabajo se centra en la aplicación de técnicas de segunda generación, dado que las de tercera y cuarta generación se encuentran aún en proceso de desarrollo (Niedringhaus et al., 2011). Como se ha comentado, existen distintas plataformas desarrolladas para la aplicación de las técnicas de NGS y la elección de las plataformas de secuenciación depende de cuales sean los objetivos. Las principales plataformas comerciales en la segunda generación de NGS son: Roche/454®, Illumina/Solexa® y Applied Biosystems/SOLID®. Las diferentes plataformas se distinguen por las tecnologías utilizadas para secuenciar: las plataformas de Roche e Illumina utilizan la llamada secuenciación por síntesis (Niedringhaus et al., 2011), mientras que SOLID®, de Applied Biosystems aplica un método de secuenciación por ligación (McKernan et al, 2009). A su vez, mientras Roche usa el método de emulsión por PCR para amplificar las moléculas de ADN, Illumina utiliza la amplificación en fase sólida, diferenciándose ambas también en las tecnologías de detección de fluorescencia (Metzker, 2010). A la hora de escoger una u otra plataforma de secuenciación es importante conocer las diferencias en los tipos de datos que generan, como por ejemplo, distintas puntuaciones de calidad de las bases, diferencias en la precisión de la secuenciación y en el largo de las secuencias, entre otras.

En este trabajo se utilizó la plataforma Solexa® de Illumina, en la cual, los oligonucleótidos de secuencia complementaria a los adaptadores de los fragmentos de ADN de las librerías, se encuentran unidos covalentemente a la celda de flujo del secuenciador. Los fragmentos de ADN se inmovilizan en la celda de flujo al hibridar con los oligonucleótidos, luego se unen separados entre sí permitiendo que ocurra la formación de los clusters de moléculas de ADN idénticas. La formación de clusters se produce porque cada molécula de ADN hibrida a través de su extremo libre, con el segundo adaptador, a un nuevo oligo vecino unido a la celda, formando lo que se llaman puentes de amplificación (Figura 3). Una vez formado los mismos, una polimerasa se encarga de copiar la hebra, la cual es clivada y desnaturalizada (Figura 3). Este proceso se repite muchas veces hasta formar los clusters, los cuales cuentan con aproximadamente 1000 moléculas idénticas de ADN simple hebra. En cada celda de flujo es posible formar más de 150 millones de clusters. Es importante conocer este proceso, ya que durante el mismo es posible que se introduzcan errores, siendo fundamental su consideración para la correcta detección de SNPs. A continuación, el chip se transfiere a la

plataforma de fluidos de Illumina para la secuenciación propiamente dicha. La secuenciación se inicia con la hibridación de un cebador complementario al adaptador en el cual se va a anclar la polimerasa para iniciar la elongación, la cual adiciona de uno en uno, nucleótidos marcados fluorescentemente con 4 colores diferentes. Luego de agregar cada nucleótido, se cumple un paso de lavado, el cual es seguido del registro fluorescente por parte del detector. Para finalizar, tanto el terminador como el marcador fluorescente son removidos, cerrando el ciclo con un nuevo lavado. Estos pasos se repiten numerosas veces y cada secuencia de pasos es denominada ciclo. Este sistema es capaz de producir lecturas de un largo de 100 pb ya que para obtener mayores tamaños sería necesario incrementar el número de ciclos, existiendo una mayor probabilidad de producir un desfasaje en la elongación de las hebras de ADN (Erlich et al., 2010). Si esto ocurre, la presencia de una población heterogénea de lecturas dentro de un cluster produce ruido de fondo, el cual disminuye la pureza de la señal y por ende, la precisión. Esta limitación no es tan significativa en la plataforma de 454® de Roche, donde se obtienen lecturas de largos mayores entre 400 y 500 pb (Knierim et al., 2011).

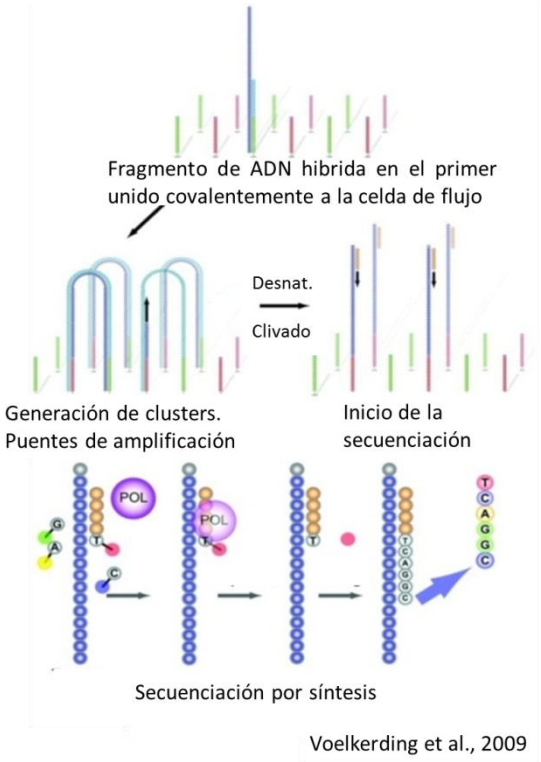



Figura 3. Esquema representativo del proceso de secuenciación de ADN mediante la plataforma Illumina/Solexa. Se detalla la unión del fragmento de ADN a la celda de flujo, la formación de los puentes de amplificación y los clusters correspondientes y por último, la secuenciación mediante la adición de nucleótidos marcados fluorescentemente.

El error más común producido por esta metodología es la sustitución de una base nucleotídica por otra. Esto es aún más frecuente cuando el nucleótido anteriormente adicionado es una Guanina (Dohm et al., 2008). También se ha registrado una mayor pérdida de fidelidad al secuenciar regiones ricas en repetidos AT y GC, probablemente debido a errores durante la amplificación en la construcción de los clusters (Dohm et al., 2008; Hillier et al., 2008; Harismendy et al., 2009). A pesar de que la plataforma Illumina presenta limitaciones, la cantidad de datos que genera y la accesibilidad en relación a los costos, en comparación con otras plataformas, hace que sea una de las tecnologías más utilizadas actualmente. Es una tecnología que ha sido exitosamente aplicada en la secuenciación y genotipado de genomas de grandes tamaños, como por ejemplo trigo y maíz (Akhunov et al., 2009, Elshire et al., 2011).

**Identificación de SNPs – Análisis bioinformático**

Una vez generados los datos de secuenciación, el desarrollo en el área de la tecnología y la bioinformática permite disponer de las herramientas necesarias para identificar posibles marcadores a partir de las secuencias crudas provenientes del secuenciador. Para analizar estos datos es necesario contar con procesadores que permitan almacenar y procesar Giga a Terabytes de datos. No existe una metodología estándar para el análisis de secuencias, variando con la especie en estudio, así como también con los métodos de construcción de librerías y de secuenciación que se utilicen (Nielsen et al., 2011).

Un proceso factible para el análisis de los datos es la división en dos etapas. La primera, que comprende el procesamiento de secuencias y el alineamiento las mismas (ya sea contra un genoma de referencia o entre las mismas secuencias) y la segunda, culminaría con la identificación de los SNPs y el genotipado (Yang et al., 2010). El objetivo del análisis es alinear las secuencias provenientes de la re-secuenciación del genoma, con el fin de identificar posibles variaciones entre los genotipos que han sido re-secuenciados. Estas variaciones pueden identificarse al ser comparadas contra un genoma de referencia (Deschamps et al., 2010; Wu et al., 2010; Varala et al., 2011; Hendre et al., 2012) o también en ausencia de este, siendo más complejo el análisis en este último caso. La ausencia de genoma de referencia introduce una dificultad en el análisis y para sortearla se han desarrollado diversas estrategias. Una de ellas consiste en la construcción de un genoma *de*

*novo* mediante el ensamblado de secuencias obtenidas con la plataforma 454 de Roche (Hyten et al., 2010c; You et al., 2011). Una segunda estrategia se basa en el alineamiento de las secuencias de los distintos genotipos entre sí, con el fin de identificar las variaciones genéticas deseadas (Ratan et al., 2010; Elshire et al., 2011; Poland et al., 2012a). Una vez realizado el alineamiento, el siguiente paso, es la identificación de posibles candidatos a SNPs y la obtención de los genotipos correspondientes a las distintas muestras. Finalmente es necesario filtrar falsos SNPs provenientes de errores de secuenciación, secuencias homeólogas o parálogas. Existen diversas herramientas disponibles para la realización de estos análisis, algunos de los paquetes informáticos son: Fastxtoolkit (Blankenberg et al., 2010a y 2010b), SAMtools (Li et al. 2009), vcftools (Danecek et al., 2011), GATK (McKenna et al., 2010; DePristo et al., 2011) y varscan (Koboldt et al., 2009).

## 1.4. SNPs EN EL ANÁLISIS DE ESTRUCTURA DE POBLACIONES Y ANÁLISIS DE DESEQUILIBRIO DE LIGAMIENTO

Los polimorfismos entre las distintas variedades de un cultivo son indicativos de la diversidad genética presente entre los individuos (Zhang et al., 2011). En poblaciones menos diversas se obtiene un menor número de estos marcadores, dado que los individuos poseen un genoma más similar, situación que ocurre en análisis de poblaciones de mapeo biparentales (Hyten et al., 2010a). El conocimiento de la diversidad genética y de la estructura poblacional entre los individuos es muy importante a la hora de escoger una estrategia en la mejora genética de un cultivo (Chen et al., 2012). En este sentido, la posibilidad de encontrar genes que expresen características de interés en el cultivo, depende en gran medida de la variabilidad genética que exista dentro de los individuos de la población (Flint-García et al., 2005). La estructura de la población está definida por la formación de subpoblaciones que poseen mayores valores de parentesco genético entre sí por haber co-evolucionado separadamente. Estudiar la estructura de la población nos permite conocer la historia de un programa de mejoramiento, que incluye los ancestros pre-domesticados, el sistema de mejoramiento usado y la complejidad de las prácticas de mejoramiento aplicadas (Xie et al., 2008). El genotipado con marcadores en el conjunto de individuos a estudiar nos permite construir matrices de distancias genéticas, las cuales son valores de pares de individuos que indican el grado de cercanía genética entre dicho par (Guo et al., 2011).

Para poder analizar los datos generados existen diferentes métodos que facilitan su interpretación. Los análisis más comúnmente aplicados son análisis de componentes principales (PCA, "Principal Component Analysis") o análisis de coordenadas principales (PCoA, "Principal Coordinate Analysis"), análisis de clúster de clasificación supervisada y no supervisada. El PCA es un análisis sencillo, que permite identificar patrones en los datos, resaltando sus diferencias y similitudes, mediante la disminución de las dimensiones de estos para facilitar su interpretación (Smith, 2002). El PCoA posee una base similar que el PCA pero está diseñado para aplicarse en caso de tener matrices de distancia genética, siendo útil para obtener información de estructura de los resultados en base a los datos genéticos obtenidos (Dreisigacker et al., 2005; Zhang et al., 2011).

La información de pedigrí, que reconstruye la historia de cada línea, también permite obtener matrices de parentesco entre los individuos. Algunos trabajos demuestran que las matrices de parentesco construidas a partir de marcadores moleculares exhiben mayor precisión a la hora de establecer estas relaciones genéticas (Almaza-Pinzón et al., 2003; Condón et al., 2008).

Por otro lado, el análisis de desequilibrio de ligamiento (LD) es comúnmente utilizado para identificar grupos de asociación entre marcadores (grupos de LD). Estos grupos permiten construir haplotipos o grupos de marcadores que segregan juntos en las líneas en estudio, eliminando del análisis datos redundantes y generando marcadores poli-alélicos (Akey, 2001; Wall y Pritchard, 2003). Esto es posible debido a que marcadores cercanos en el genoma tienden a segregar juntos durante la meiosis, cuando se dan los procesos de recombinación entre los cromosomas (Flint-García et al., 2003; Gupta et al., 2005). Existen distintas formas de calcular este valor de LD entre pares de marcadores, los más utilizados son $r^2$ y D'. D' solo estima diferencias de recombinación, mientras que $r^2$ además, toma en cuenta la historia de mutaciones (Jorde, 2000; Ardlie et al., 2002). En un cultivo con alto desequilibrio de ligamiento como el trigo (Sommers et al., 2007) se esperaría que se formaran diversos grupos de ligamiento si se cuenta con un buen número de marcadores moleculares ordenados a lo largo del genoma.

## 1.5. ASOCIACIONES ENTRE GENOTIPO Y FENOTIPO

La expresión de un fenotipo en un individuo depende tanto del componente genético como de los factores ambientales en los cuales está expuesto. Los factores ambientales incluyen

características intrínsecas de un determinado sitio, variables climáticas, de manejo del cultivo, entre otras. Los factores genéticos básicamente están determinados por el ADN de cada una de las plantas. Si bien esta es una subdivisión ampliamente aceptada, ambos factores no son independientes, y en función de la característica fenotípica que se considere, pueden estar más o menos relacionados entre sí (Bnejdi y Gazzah, 2010; Chenu et al., 2011). La heredabilidad es un factor característico de las variables fenotípicas que permite estimar el grado de interacción entre los dos factores en la expresión del fenotipo de interés. En definitiva, nos indica cuánto de la variable está explicado por la genética de la planta y cuanto es debido al ambiente, siendo las características con alta heredabilidad menos influenciadas por este (Holland et al., 2003; Wray y Visscher et al., 2008). Algunos de los caracteres son de herencia simple o cualitativa, en los cuales intervienen pocos genes de efectos aditivos (el efecto del heterocigoto tiene un efecto promedio entre ambos alelos). Estos caracteres han sido ampliamente estudiados y no constituyen actualmente un objetivo en la mejora genética de cultivos. Las características de mayor importancia agronómica, como ser rendimiento, calidad y resistencia a enfermedades, presentan en general herencia compleja o cuantitativa. Lo anterior refiere a que están determinadas por numerosos genes que contribuyen diferencialmente en la determinación del fenotipo (Mackay et al., 2009). Lo que mayoritariamente ocurre es que existen pocos genes de efectos mayores que son fáciles de detectar y varios genes de efectos menores, que tienen acción pleiotrópica sobre el carácter estudiado (Robertson, 1967; Kearsey y Farquhar, 1998). El efecto genético no es únicamente aditivo, existe un componente que es no aditivo en el cual intervienen otros tipos de acciones genéticas, por ejemplo, dominancia completa o incompleta y epistasis, que complejizan aún más los análisis (Carey, 2000). Por lo tanto, los estudios que buscan detectar posibles asociaciones entre el genotipo y el fenotipo deben tener en cuenta esta diversidad de factores a la hora de identificar genes involucrados en diferentes características o identificar genotipos que posean el fenotipo deseado.

El análisis de locus de carácter cuantitativo (QTL, "Quantitative Trait Locus") comenzó a cobrar importancia entre la década del 80 y el 90 con los primeros estudios llevados a cabo en tomate y maíz (Paterson et al., 1988, Stuber et al., 1992). Este análisis posee como objetivo principal identificar el número, la posición y el efecto, de los genes involucrados en un carácter de herencia cuantitativa (Lande y Thompson, 1990). El QTL se define como aquella región, en alguna parte del genoma, que está involucrada en la expresión del carácter cuantitativo (Collard et al., 2005). El análisis de QTL se lleva a cabo en poblaciones de

mapeo biparentales que permitan identificar marcadores moleculares que segreguen junto a los QTL de interés, es decir, que se encuentren en mayor proporción en los individuos con el fenotipo de interés (Huang et al., 2011). Estos marcadores tienen una dirección en el genoma y permiten asignar una localización aproximada a los QTL. Esta metodología fue ampliamente aplicada a distintos cultivos, permitiendo identificar regiones genómicas involucradas en la expresión de caracteres complejos (Andaya y Mackill, 2003; Kirigwi et al., 2007; Cuesta-Marcos et al., 2009). Estos trabajos no solo permitieron avanzar en el conocimiento de los mecanismos genéticos subyacentes, sino que también aportaron una herramienta en la selección y manipulación de los caracteres (Tanksley y Nelson, 1996). Una de las limitaciones de este tipo de análisis, es la necesidad de contar con poblaciones segregantes y balanceadas para la obtención de una base genética similar en la población, lo cual facilita la detección de diferencias debido al carácter en estudio (Huang et al., 2011). Otra limitante es la necesidad de contar con un mapa de ligamiento para la población a analizar (Hori et al., 2003). A su vez, para su aplicación en el marco de un programa de mejoramiento, los QTL conocidos deben ser previamente validados en la población local (Parisseaux y Bernardo, 2004; Ordas et al, 2010; Castro et al., 2012). Una vez realizada esta validación, la selección de los QTL con marcadores moleculares implica pasos de amplificación y corrida de geles en cientos a miles de individuos para múltiples QTL.

Como alternativa al análisis de QTL existe el Mapeo asociativo, el cual se centra en la búsqueda de asociaciones de marcadores con fenotipos de interés, basándose en el concepto de LD entre genes y marcadores (Myles et al., 2009). Estos análisis intentan eliminar la necesidad de desarrollar poblaciones de mapeo para la identificación de asociaciones genéticas (Filt-Garcia et al. 2003; Varshney et al. 2005), trabajando directamente con los individuos que componen el programa de mejoramiento. Existen diferentes variables que afectan este tipo de análisis como ser el grado de LD, la estructura de la población (Pitchard et al., 2000), la deriva génica y selección (Filnt-García et al., 2003). Debido a lo anterior, previo a la realización de estos análisis es necesario conocer las características genéticas del material en estudio, así como la estructura de la población, ya que de esto dependerá en gran medida la cantidad de marcadores y el número de individuos necesarios para los análisis. Recientemente, con el desarrollo de nuevas tecnologías de obtención de marcadores es posible obtener altas densidades de marcadores distribuidas a lo largo del genoma, lo cual permite realizar estudios de asociación conocidos como “Genome Wide Asociation Mapping” (GW-AM). Este tipo de estudios permiten realizar simultáneamente asociaciones

marcador – fenotipo en todo el genoma mediante un análisis de modelos mixtos (von Zitzewitz et al., 2011).

La selección genómica (SG) por su parte, es una nueva metodología de selección por marcadores propuesta por Meuwissen et al. (2001), en la cual se seleccionan los mejores individuos basándose en el valor de cría genético estimado (GEBV). Esta terminología, poco difundida hasta el momento en el mundo vegetal, pero inicialmente utilizada por medio de selección con matrices de pedigrí, es ampliamente utilizada en animales, donde cada animal posee un valor de cría genético (Hayes et al, 2009; Nayaka y Isobe, 2012). El principio de este análisis es realizar predicciones fenotípicas en base al genotipo de los individuos. Para esto es necesario entrenar un modelo con información genética y fenotípica de un conjunto de individuos que formen parte del programa de mejoramiento. Este modelo entrenado es capaz de predecir el comportamiento fenotípico de variedades que no presentan información fenotípica (Mewissen et al., 2001; Heffner et al., 2011). A diferencia del análisis de QTL, SG permite analizar toda la varianza genética de cada individuo mediante la suma de los efectos de los alelos de GEBV (Heffner et al., 2009), por lo que se espera capturar regiones genómicas con menor efecto en el fenotipo, las cuales se enmascaran en un análisis de QTL clásico (Goddard y Hayes, 2009). Existen diversos estudios en plantas que demuestran el éxito de la selección genómica (de los Campos et al., 2009; Heffner et al., 2009; Crossa et al., 2010; Iwata y Jannink, 2011; Resende et al., 2011). El paso siguiente sería comenzar a aplicar esta metodología en un programa de mejoramiento.

## 1.6. SELECCIÓN GENÓMICA Y MEJORAMIENTO VEGETAL

La SG podría constituirse como una herramienta clave en los programas de mejoramiento tanto vegetal como animal, permitiendo la predicción de caracteres fenotípicos complejos. Actualmente, para poder seleccionar líneas experimentales de trigo con alto potencial de rendimiento y niveles aceptables de tolerancia a estrés biótico y abiótico, es necesario esperar al desarrollo de la planta bajo diferentes condiciones de crecimiento y evaluar el comportamiento de los distintos genotipos en estas condiciones (Heffner et al, 2010). Un programa de mejoramiento se inicia con un gran número de cruzamientos iniciales, el cual dependerá de las capacidades de los programas de mejoramiento. En el caso del programa de mejoramiento genético de trigo de INIA La Estanzuela, el proceso se inicia con unos 600

cruzamientos. La generación F1 pasa inmediatamente a F2 dado que en F1 no existen variaciones, por ser todos los individuos genéticamente idénticos. En la población F2 es clave la correcta selección de los genotipos, ya que es la generación en la cual ocurren las mayores variaciones entre los individuos. En esta generación se selecciona por caracteres simples, los cuales poseen en general, altos valores de heredabilidad (por ej.: resistencia a algunas enfermedades, altura de planta, tamaño de grano, entre otros). A continuación, los diferentes genotipos se autofecundan hasta la generación F6 y en cada una de las generaciones se realiza selección por fenotipo. Recién a nivel de F5 y F6 se comienza a seleccionar por rendimiento, dado que se cuenta con un menor número de individuos, lo cual permite realizar ensayos a mayor escala. Finalmente, se cuenta con un grupo de individuos preliminares que serán seleccionados y evaluados por rendimiento, obteniendo los genotipos finales que son evaluados en distintos ambientes y con las repeticiones correspondientes. Los individuos con mejores rendimientos y que no presenten ninguna característica indeseable son seleccionados y denominados "elite". Las variedades elite son evaluadas durante 3 años en un programa de evaluación antes de poder ser liberadas. Todo este proceso lleva entre 10 y 15 años. (M. Quincke, com pers, 2012).

Con el objetivo de acortar estos tiempos y aumentar la eficiencia en combinación con métodos de avance generacional (haploides duplicados, descendencia por semilla única en autógamas y avance generacional fuera de época, entre otros), la SG busca inferir esta información, basándose en el ADN del individuo (Heffner et al., 2011). Uno de los objetivos más interesantes, dentro de un programa de mejoramiento, es poder realizar SG en la generación F2, seleccionando desde el comienzo los genotipos de interés. Otro desafío de la SG es predecir los cruzamientos que generan individuos con mayores potenciales (Endelman et al., 2011). Para realizar estas predicciones son necesarias diversas herramientas que se encuentran aún en desarrollo. Las principales son:

- Amplia y densa cobertura de marcadores a lo largo de todo el genoma. Es importante enfatizar que esta herramienta debe ser accesible (en referencia a su simplicidad y costos) de ser aplicada en un programa de mejoramiento (Elshire et al., 2011, Poland et al., 2012a).
- Métodos de fenotipado de alta precisión (revisado por Cabrera-Bosquet et al., 2012).
- Desarrollo de modelos estadísticos adecuados que permitan realizar los análisis para calcular los valores de cría estimados (Meuwissen et al., 2001; Crossa et al., 2010).

## 1.7. OBJETIVOS

### 1.7.1. Objetivo general

Ajustar y validar una metodología de identificación de SNPs para su aplicación en programas de mejoramiento genético de trigo.

### 1.7.2. Objetivos específicos

I. Ajustar y validar una metodología de análisis bioinformático para la identificación de marcadores, mediante el análisis 384 genotipos de trigo de los programas de mejoramiento genético de INIA Uruguay e INIA Chile.

II. Validar la metodología de selección genómica como herramienta de selección en el mejoramiento genético de trigo.

# 2. METODOLOGÍA

## 2.1. GENOTIPADO

Los trabajos se llevaron a cabo a partir de una selección 384 materiales pertenecientes a los programas de mejoramiento de INIA Chile e INIA Uruguay, de los cuales 186 genotipos pertenecen al programa de mejoramiento de Uruguay y 198 al de Chile. Para la extracción de ADN se utilizaron 100 mg de tejido vegetal, empleando el kit de extracción DNeasy Plant Maxi Kit (Qiagen), asegurando la pureza necesaria para el análisis. La concentración y pureza del ADN en las muestras se determinó por espectrofotometría y mediante la utilización de geles de agarosa al 1 %.

Para la construcción de las librerías se prepararon diluciones a una concentración de 20 ng/µl de ADN, en un volumen total de 15 µl de buffer de dilución (Qiagen), las cuales se enviaron al laboratorio de secuenciación del Institut de Biologie Intégrative et des Systèmes (IBIS) de la Université Laval (Quebec, Canadá). Las librerías se construyeron según el protocolo propuesto por Elshire et al. (2011), modificado con enzimas de restricción PstI/MspI (Poland et al., 2012a). Posteriormente las muestras fueron enviadas a la Universidad de McGill de Montreal para su secuenciación en la plataforma Illumina/Solexa®.

## 2.2. IDENTIFICACIÓN DE MARCADORES

Los datos de la secuenciación fueron recibidos en un archivo crudo proveniente del secuenciador, en formato Fastq. En un principio se analizó la calidad de las secuencias y la distribución nucleotídica, utilizando el software Galaxy (http://galaxy.psu.edu/). A continuación, se estudió la distribución de las secuencias en las 96 muestras de cada línea de la celda de flujo, para constatar si todas las muestras presentaban lecturas y evaluar la distribución entre las distintas muestras.

Se identificaron SNPs mediante dos metodologías diferentes de análisis de datos. La primera metodología empleada fue desarrollada en el Departamento de Agricultura de EE.UU. (USDA, www.maizegenetics.net) y se denominó Pipeline de Tassel (PipT), permite identificar variaciones en los fragmentos secuenciados sin la necesidad de contar con un

genoma de referencia. La segunda metodología, desarrollada para el cultivo de soja en la Universidad de Laval, Quebec, Canadá (sin publicar), se diferencia de la anterior por contrastar los individuos contra un genoma de referencia. En este trabajo se realizó una adaptación de la misma para poder ser aplicada en el cultivo de trigo, el cual carece (por el momento) de un genoma de referencia. Ambas metodologías se iniciaron con la identificación de los códigos de barras y adaptadores asignados a las distintas lecturas de cada muestra (llamaremos lecturas a las secuencias provenientes del secuenciador Illumina), descartando lecturas que no cumplieron con los requisitos de calidad. Para trabajar con las lecturas fue necesario eliminar los códigos de barras que no forman parte real de la secuencia del genoma. A partir de este punto, ambas metodologías operan de un modo diferente con el fin de identificar las variaciones en una sola base. Una vez identificadas estas variaciones, se filtraron SNPs para eliminar la mayor cantidad de falsos marcadores. Ambas metodologías se detallan a continuación.

### 2.2.1. Pipeline de Tassel

A continuación se describe el análisis de los datos obtenidos en la secuenciación, mediante la aplicación del protocolo obtenido en USDA, Kansas, EE.UU. En el primer paso se seleccionaron las lecturas válidas y se contabilizó el número de veces que se repitió cada lectura. Las lecturas seleccionadas fueron aquellas que contaron con el código de barras seguido del sitio de restricción de la enzima y que no presentaron ninguna N (bases no identificadas durante las secuenciación en esa posición) en las primeras 64 pb. Estas lecturas fueron cortadas en fragmentos de 64 pb y colocadas en un nuevo archivo, el cuál contiene secuencias que se denominan tags. Estas secuencias representan un conjunto de lecturas idénticas, identificándose la cantidad de secuencias que conforman este grupo.

En el siguiente paso se identificaron las muestras que presentaron cada lectura correspondiente. Las muestras se procesaron de la misma manera detallada en el paso anterior, obteniéndose un archivo que contiene las distintas tags en las filas y los individuos (muestras) en las columnas. Cuando la tag correspondiente estaba presente en la muestra se asignó un valor en la matriz que corresponde al número de lecturas que formaban dicha tag en esa muestra y por el contrario, si dicha tag no estaba presente en la muestra, se asignó un valor de 0.

Es importante aclarar que para cumplir con los dos pasos anteriores es necesario construir un archivo en el cual se identifica a cada muestra con su código de barras, este archivo se conoce como "Key". Paralelamente, es necesario también, proporcionar información sobre que enzima de restricción fue utilizada en la construcción de las librerías de GBS (en este trabajo fueron utilizadas las enzimas PstI-MspI). Al finalizar este paso se cuenta con la información organizada para poder avanzar en la realización de alineamientos de a pares entre las tags para poder identificar los posibles SNPs. Una vez identificados los SNPs se colocan en formato HapMap para poder continuar el análisis.

### 2.2.2. Pipeline con genoma de referencia

En esta segunda metodología de análisis, desarrollada en la Universidad Laval, Quebec, Canadá para el genoma de soja (sin publicar), el análisis se inició con la separación de secuencias por código de barras comunes. Se generó un archivo fastq para cada muestra secuenciada. A continuación, las muestras fueron cortadas en fragmentos para eliminar el código de barras y el adaptador; todas las lecturas se redujeron a un tamaño no mayor de 104 pb, eliminando los fragmentos conformados por menos de 35 pb. Como se ha mencionado, esta metodología requiere de un genoma de referencia, por lo que se utilizó un ensamblado del genoma de trigo, disponible en (bio-sanger.bio.bris.ac.uk). El ensamblado fue construido secuenciando la variedad *Triticum aestivum* cv. Chinese Spring en la plataforma Roche/454® con cobertura 5x.

A continuación, se alinearon una a una las distintas muestras "contra" el genoma de referencia, utilizando el algoritmo Burrows-Wheeler del paquete bwa (Li y Durbin, 2009).

Como resultado de este alineamiento se obtiene, para cada una de las muestras, un archivo que contiene información del alineamiento con el genoma de referencia. Para la identificación de SNPs en el conjunto de muestras, se utilizó el paquete de herramientas conocido como SAMtools (Li, 2009) y se obtuvo un único archivo (con extensión bcf) que contiene los candidatos a SNPs por muestra. El archivo obtenido en el paso anterior necesita ser filtrado dado que contiene muchos SNPs, pero la información es incompleta. Para filtrar este archivo se utilizó el paquete de herramientas llamado vcftools (Danecek et al., 2011). Luego de realizada la filtración, se obtuvo un archivo final con extensión vcf que puede ser

convertido en otros formatos (por ejemplo, plink), dependiendo del tipo de análisis a desarrollar.

## 2.3. EVALUACIÓN DE LAS METODOLOGÍAS UTILIZADAS EN LA IDENTIFICACIÓN DE SNPs (PIPELINES)

Una vez obtenidos y filtrados los SNPs, se compararon ambas metodologías tomando como indicadores el análisis de la cantidad de datos faltantes por SNP y por muestra, el contenido de heterocigotos y la frecuencia alélica. Para llevar a cabo estos análisis se utilizó el software R (http://cran.r-project.org/).

## 2.4. VALIDACIÓN DE SNPs

Para confirmar si los SNPs identificados son verdaderas variaciones entre individuos ubicados en una misma región genómica, se seleccionaron distintos métodos de validación de los mismos, los cuales se interrelacionan entre si y se describen a continuación.

### 2.4.1. Comparación entre matrices genéticas con sub-grupos de datos

En este procedimiento de validación, el total de SNPs candidatos identificados se dividieron al azar formando dos subgrupos, y se calcularon matrices de relaciones aditivas (A) entre los individuos para cada grupo. La matriz A se calcula de la siguiente forma (VanRaden, 2008):

$A = WW'/c \qquad W_{im} = X_{im} + (1 - 2p_m)$

$c = 2\sum_m p_m (1 - p_m)$

$p_m$ es la frecuencia del alelo 1 en el marcador m, X es la matriz de marcadores codificada como -1, 0, 1; m: marcadores; i: individuos.

Como se cuenta con una gran proporción de datos perdidos, debido a las características de la metodología de genotipado utilizada (GBS), se imputaron los datos faltantes (Poland et al., 2012a) utilizando el paquete rrBLUP (Endelman et al., 2011) disponibles en el software R. Para realizar las comparaciones, mediante la correlación de Pearson con el test de Mantel (paquete vegan del software R), se utilizaron los datos con y sin imputación (la imputación se realizó por separado en cada subgrupo de los SNPs).

Este método podría considerarse de screening para la validación, dado que, si los SNPs son verdaderas variaciones en el genoma, se cuenta con suficientes marcadores para esperar obtener matrices muy parecidas al calcularlas con la mitad de los SNPs identificados.
Finalmente, se calcula las matrices de relaciones aditivas con el total de los SNPs obtenidos en PipT y PipRef, por separado, y se comparan ambas matrices entre si.

### 2.4.2. Análisis de la estructura poblacional

A continuación se propone validar los SNPs mediante el análisis del relacionamiento genético entre los individuos. Por lo tanto, a partir de la matriz de relaciones genéticas aditivas (calculada con el conjunto de marcadores en el paso anterior) se realizó un análisis de coordenadas principales, que permite interpretar los datos disminuyendo sus dimensiones a 2 vectores principales. Este análisis se realizó con el objetivo de identificar sub-poblaciones dentro de los 382 individuos. Como herramienta informática en este caso se utilizó en paquete ape disponible en R (Paradis et al., 2004).

### 2.4.3. Comparación entre matriz de coancestría y de parentesco con marcadores

En este método de validación de SNPs, se construyó una matriz de coancestría (COP), utilizando los datos de pedigrí disponible para 177 de las 182 líneas uruguayas genotipadas. La matriz de relaciones genéticas (K) se calculó utilizando el paquete Efficient Mixed-Model Association (EMMA) de R (Kang et al., 2008), de la forma:

$K=\sum_h L_h*W*x_h$

$L_h$ es la matriz de marcadores codificada 0 y 1; W es una matriz que asigna peso diferencial a cada SNP (en este caso es un matriz diagonal de 1); $x_h$ es el efecto aleatorio del alelo.

Una vez definidas ambas matrices se compararon entre sí mediante la evaluación de la correlación de Pearson. La matriz de coancestría se construye utilizando los software ORDARCH y GENARCH (F. Condón, INIA, Uruguay) y el paquete pedigrí disponible en el software R (Coster, 2011). El objetivo de esta metodología es transformar los datos de marcadores moleculares y de pedigrí en datos de relaciones genéticas entre individuos, para poder comparar ambas informaciones.

### 2.4.4. Mapeo de SNPs

Otra forma de validar los SNPs es evaluar su distribución en el genoma de trigo, con el objetivo de observar cómo se distribuyen los marcadores en los distintos cromosomas y genomios de trigo. En esta metodología, se utilizaron las 2 bases de datos de trigo para mapear las tags que contienen los SNPs identificados. Una de las bases de datos disponibles fue generada por Poland et al. (2012a) y la otra pertenece al consorcio de trigo y está disponible en http://urgi.versailles.inra.fr/. Como herramienta informática para el mapeo se utilizó el paquete standalone blast-ncbi (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

La base de datos generados Poland et al. (2012a) está compuesta de tags secuenciadas por la metodología de GBS, en la cual se utilizó una población segregante de dobles haploides y se construyó un mapa de bins que permite ordenar las tags dentro de cada cromosoma con ayuda de otros marcadores moleculares. Por otra parte la base de datos de URGI únicamente nos permite asignar en que brazo del cromosoma se encuentra cada SNP.

### 2.4.5. Análisis de desequilibrio de ligamiento

Esta metodología de validación de SNPs analiza la existencia de desequilibrio de ligamiento utilizando los resultados obtenidos en el punto anterior. A partir de los SNPs mapeados se analizó el desequilibrio de ligamiento entre los marcadores con los genotipos de los 382 individuos. Se obtuvo el $r^2$ para todos lo pares de marcadores, utilizando como herramienta informática el paquete trio (Schwender et al., 2012), disponible en el software R. A continuación utilizando el paquete LDheatmap (Shin et al., 2006) se evaluó la presencia o ausencia de desequilibrio de ligamiento entre marcadores. El objetivo de este análisis es determinar si los marcadores, ordenados en el cromosoma, forman grupos de asociación a lo largo del cromosoma.

### 2.4.6. Predicciones fenotípicas.

Con el objetivo de validar los SNPs, en su aplicación en el cálculo de predicciones fenotípicas a partir de los genotipos secuenciados, se utilizaron datos de rendimiento y peso de mil granos obtenidos a campo (para todos los individuos) en Santa Rosa, Chile. El ensayo fue sembrado bajo un diseño de alfa látice, con 400 genotipos distribuidos en 20 bloques de 20 tratamientos por bloque, con dos repeticiones por tratamiento. El mismo ensayo se realizó bajo 2 condiciones hídrica contrastantes, riego y secano.

Debido a la existencia de variación espacial en los ensayos a campo (correlación entre los residuales de las variables fenotípicas), se evaluaron distintos modelos con el objetivo de lograr el mejor ajuste de la medias. Se planteó la evaluación de 4 modelos distintos: Látice Simple (LS, diseño planteado a campo), filas y columnas (FC), bloques completos al azar con medias móviles como covariable (BCA_MVNG) y una regresión lineal con la medias móviles como covariable (MVNG). Los datos se analizaron utilizando los paquetes lme4 y mvngGrAd disponibles en el software R.

Los primeros tres modelos evaluados se plantearon como modelos mixtos de la siguiente forma:

$y = Xb + Zu + e$

en el cual X es la matriz de diseño de efectos fijos, b es el vector de los efectos fijos (BLUE), Z es la matriz de diseño de los efectos aleatorios, u es el vector de efectos aleatorios (BLUP) y e comprende los errores experimentales del modelo.

En el caso de los modelos LS y FC, X está compuesta por las repeticiones. Mientras que para BCA_MVNG además de las repeticiones se incluye la covariable ($x_i$) de las medias móviles en los efectos fijos del modelo. En el caso de LS los efectos aleatorios son los tratamientos y bloques anidados en las repeticiones; en el caso de FC, dichos efectos están constituidos por los tratamientos y las filas y columnas anidadas en las repeticiones y, en el caso de BCA_MVNG, por los tratamientos. El modelo MVNG considera la covariable ($x_i$) de las medias móviles como regresor en el modelo. $x_i = y_i - \bar{y}_{grid}$, donde $y_i$ es el valor fenotípico de cada observación y $\bar{y}_{grid}$ es la media fenotípica de las observaciones que se consideran dentro del grid. El grid fue construido utilizando el paquete movingGrid de R (Technow, 2011; Figura 4).

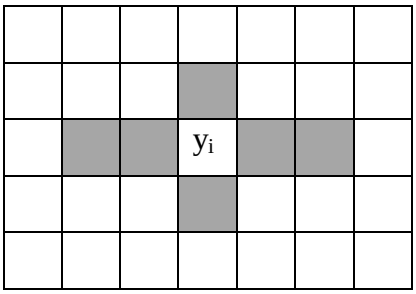



Figura 4. Los plot señalados son los que forman parte del grid con el cual se calcula la media móvil, que posteriormente se emplea para restar al valor de $y_i$ en el cálculo de la covariable.

Tabla 1. Descripción detallada de las ecuaciones que representan los 4 modelos usados para ajustar las medias fenotípicas en el campo.

| | **Modelos** |
|---|---|
| **LS** | $y = g_i + rep_j + bl(rep)ijk + e_{ijk}$ |
| **FC** | $y = g_i + rep_j + fil(rep)_{jk} + col(rep)_{jl} + e_{ijkl}$ |
| **BCA_MVNG** | $y = g_i + \beta x_j + rep_k + e_{ijk}$ |
| **MVNG** | $y = u + \beta x_i + e_i$ |

g: tratamientos, rep: repeticiones, bl(rep): bloques incompletos anidados dentro de las repeticiones, fil(rep): filas anidadas en repeticiones, col(rep): columnas anidadas en repeticiones, e: residual, x: covariable calculada como el valor fenotípico del plot menos la media fenotípica de las parcelas vecinas que forman el grid. LS: Diseño de campo, alfa látice simple; FC: modelo de análisis de filas y columnas anidadas en las repeticiones; BCA_MVNG: modelo de bloque completos con medias móviles como covariables; MVNG: modelo lineal medias móviles como covariable.

Luego de ajustar los datos se obtuvieron los BLUP (con el efecto genético de cada genotipo) en cada uno de los modelos, para realizar las predicciones fenotípicas con el paquete rrBLUP (ridge regression) disponible en R. Del set de 382 individuos, de los que se obtuvieron datos genotípicos, previo a la obtención de los BLUP, se eliminó la variedad 209 por presentar muy baja cobertura en la parcela. Los 381 genotipos se repartieron aleatoriamente en 7 sets (3 set con 55 individuos y 4 set con 54) para realizar la validación cruzada (Legarra et al., 2008). Seis sets (86% de individuos) fueron utilizados para entrenar el modelo de predicción. Una vez calculados los parámetros del modelo, se predijeron los valores de cría estimados (GEBV) para el set restante de validación. A continuación, se determinó la correlación entre los GEBV predichos por el modelo y los BLUP (anteriormente calculados) en el set de validación. Se realizan 10 aleatorizaciones diferentes; en cada aleatorización se determina la correlación intercambiando el set de validación hasta validar en los 7 sets. Para realizar las predicciones se utilizó el método de Genomic BLUP (G-BLUP, equivalente a Ridge Regresión) y G-BLUP utilizando el Kernel GAUSS que considera efectos epistáticos (Gianola y van Kaamm, 2008) y que se encuentra disponible en el paquete rrBLUP (Endelman, 2011).

# 3. RESULTADOS

## 3.1. GENOTIPADO

Los resultados obtenidos son producto de la secuenciación de 384 genotipos de trigo provenientes de los programas de mejoramiento de INIA Chile e INIA Uruguay. Las librerías se construyeron con el método de GBS y la plataforma de secuenciación utilizada fue Illumina/Solexa HiSeq 2000. Producto de la secuenciación de las muestras se obtuvieron 4 archivos provenientes de 4 líneas distintas de la celda de flujo del secuenciador. Cada archivo contiene secuencias mezcladas provenientes de 96 muestras en formato Fastq.

La calidad de los fragmentos resultó aceptable hasta el final de cada secuencia (108 pb). Si bien se observó un descenso hacia el extremo 3’ de los fragmentos, en ningún caso fue inferior de 30 de valor promedio (0.1 % de error) (Figura 5). La distribución nucleotídica de los fragmentos permitió verificar que los nucleótidos se presentaron con igual probabilidad en cada posición, en el conjunto de secuencias. Se observó además, distorsión en las primeras bases donde se encuentran los códigos de barras, seguidos por el sitio de restricción de la enzima (Figura 6).

La calidad de la secuenciación resultó similar en las distintas líneas estudiadas, a modo de ejemplo se presenta la figura 5.

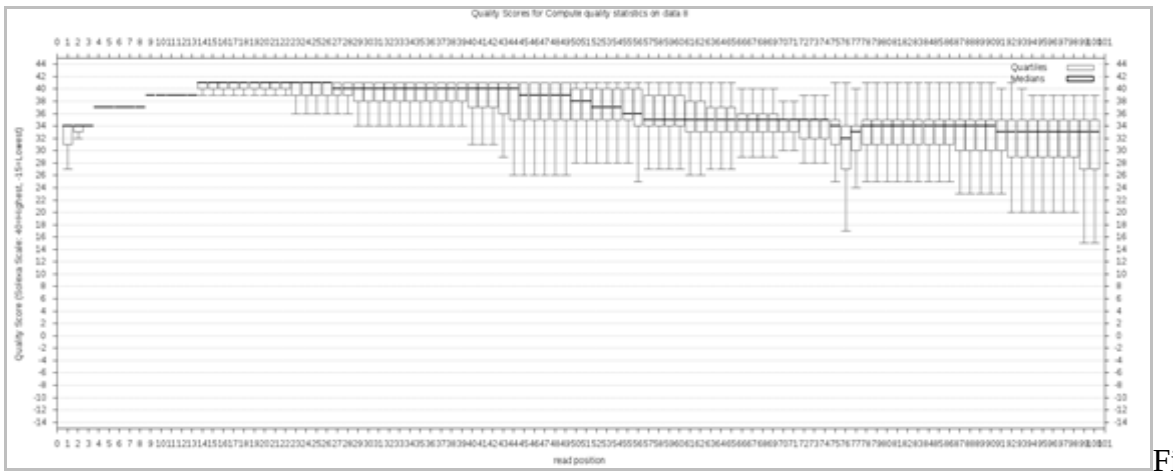



Figura 5. Distribución de la calidad en la secuenciación para cada una de las bases en el conjunto de fragmentos para una de las líneas del secuenciador.

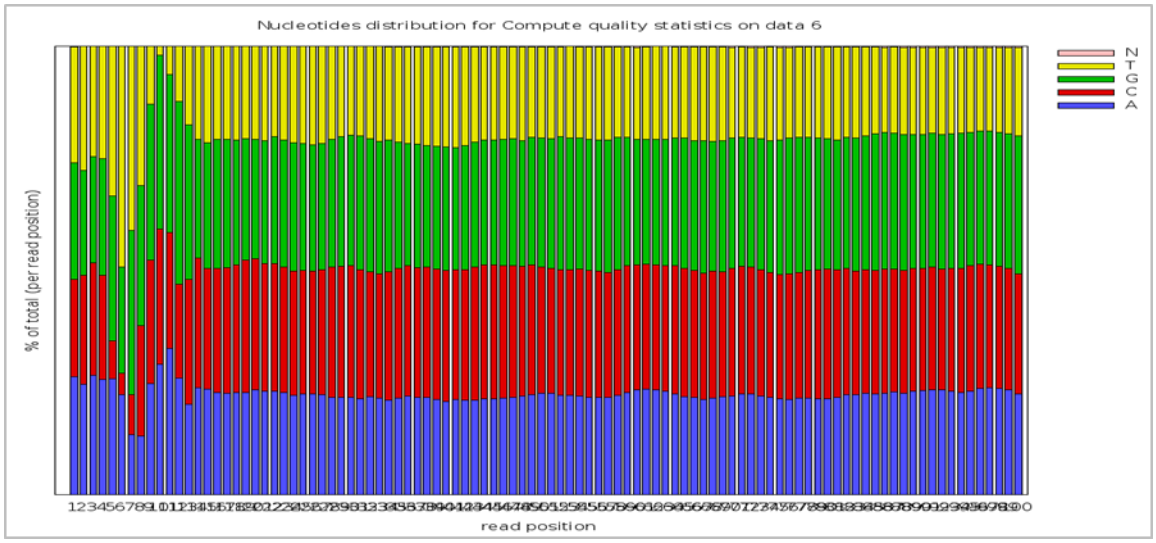



Figura 6. Distribución nucleotídica en los fragmentos en función de la posición de las bases en el fragmento de ADN secuenciado. N, no identificada; T, timina; G, guanina; C, citosina; A, adenina.

El análisis de la distribución de las lecturas por muestra fue posible al separar las secuencias por código de barras, asignando cada fragmento de ADN a la muestra de la cual proviene. Dos de las muestras secuenciadas en la línea 4 no presentaron fragmentos de ADN, lo que redujo el número total de muestras a 382 genotipos. En el resto de las muestras, la cantidad de lecturas fue aceptable, con variaciones en el número de secuencias identificadas (Tabla 2). La línea número 4 fue la que presento mayor porcentaje de fragmentos perdidos por no presentar código de barras y/o el sitio de restricción de la enzima. La media de secuencias por muestra fue similar en todas la líneas (aproximadamente 1 % de las secuencias totales).

Tabla 2. Análisis de la distribución de las secuencias por línea del secuenciador.

| | **Línea 1** | **Línea 2** | **Línea 3** | **Línea 4** |
|---|---|---|---|---|
| **Número total de fragmentos** | 165.943.288 | 175.654.571 | 153.151.244 | 181.354.392 |
| **Fragmentos descartados (%)** | 5 | 4 | 4 | 18 |
| **Fragmentos en 96 muestras** | 157.317.160 | 148.399.361 | 147.498.032 | 149.099.411 |
| **Media por muestras** | 1.638.720 | 1.769.106 | 1.536.438 | 1.586.148 |
| **Desviación estándar** | 322.839 | 630.338 | 394.675 | 613.719 |
| **Número de muestras** | 96 | 96 | 96 | 94 |

Debido a que no se detectó ninguna anormalidad que ponga en duda la integridad de la secuenciación, se procedió a la identificación de SNPs.

A continuación, se presentan los resultados de la identificación de SNPs para las dos metodologías planteadas, Pipeline con genoma de Referencia (PipRef) y Pipeline de Tassel (PipT, sin genoma de referencia). La identificación de SNPs por el método de PipRef, permitió identificar más de 200 mil SNPs, muchos de los cuales no contaban con suficiente información y otros eran falsos SNPs. Luego de establecer un mínimo de 2 lecturas para identificar el genotipo en cada muestra (de lo contrario se consideró dato perdido) y establecer un número de lecturas promedio menor a 2 y mayor a 10 (en el conjunto de las muestras), se identificaron aproximadamente 83.000 SNPs. En el conjunto de estos SNPs se identificó una alta proporción de ellos con muy baja frecuencia alélica (datos sin presentar). Los dos últimos filtros eliminaron SNPs con frecuencia alélica, menor a 0.01 y la presencia de más de 3 SNPs en una distancia de 100 pb, permitiendo identificar los 32.160 SNPs con los cuales se realizaron los subsiguientes análisis.

Por su parte el procesamiento de las secuencias según la metodología establecida en PipT permitió identificar 28.199 SNPs.

### 3.2. COMPARACIÓN ENTRE PIPELINES

Se evaluó la distribución y características de los SNPs identificados, por dos metodologías, a través del análisis de secuencias obtenidas por GBS. La baja cobertura de GBS determinó la alta proporción de datos perdidos en el conjunto de las muestras estudiadas. En el estudio de la distribución de la cantidad de datos presentes por muestra, una alta proporción de individuos (73 %) presentaron distribución normal con media aproximada del 50% de datos presentes. Sin embargo, se registró un elevado número de muestras (27 %) en las cuales se identificó menos del 30 % de los alelos (Figura 7A). Al analizar este comportamiento para cada línea de la celda de flujo (Figura 7B, Anexo 1), se identificó que dichas muestras se ubicaron mayoritariamente en la línea número 1. El análisis de la profundidad de los fragmentos de cada una de las líneas, permitió determinar le existencia de menor número de fragmentos, de profundidades (número de veces que se encuentra repetida la secuencia) entre 3 y 16, en la línea 1 con respecto a las otras tres líneas (Anexo2).

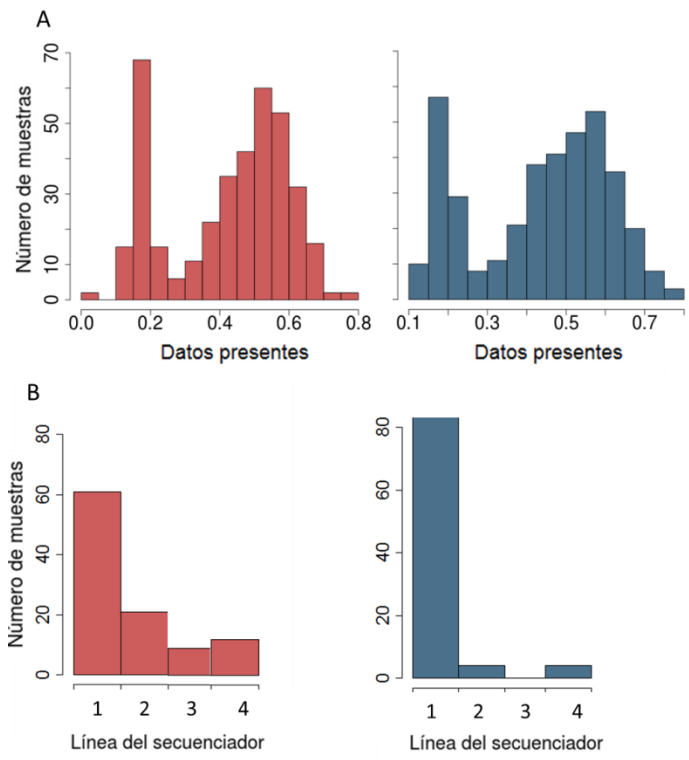



Figura 7. A) Distribución de los datos presentes en el conjunto de muestras. B) Distribución por línea del secuenciador de las muestras con más de un 70% de datos perdidos. En rojo se muestra la distribución para los SNPs obtenidos en el PipT y en azul para PipRef.

Cuando se estudió la distribución del número de SNPs según la proporción de muestras con datos presentes (Figura 8), se verificó que en ambos pipelines, la mayoría de los SNPs presentaron entre un 50 y un 70 % de datos perdidos. No se registraron SNPs con más de un 80 % de datos perdidos, ya que éste fue un criterio de filtro de falsos SNPs. Como se ha mencionado anteriormente, en ambos pipelines se filtraron los SNPs que presentaban una frecuencia alélica menor a 0.01. Se observaron diferencias entre ambas metodologías, en la proporción de SNPs con baja frecuencia alélica. En PipRef se registraron 8.800 SNPs con frecuencias alélicas entre 0.01 y 0.05, mientras que en PipT se registraron 913 SNPs en dicho intervalo de frecuencias (Figura. 8B).

Al analizar la proporción de SNPs heterocigotos en cada una de las muestras se observaron diferencias sustanciales entre ambas metodologías. En el método de PipT el 89 % de las muestras presentaron baja proporción de heterocigotos (<2 %) y ninguna muestra con más de 10%. Por lo contrario, utilizando el PipRef se observó una mayor proporción de SNPs heterocigotos (75 % de la muestras con más del 2% de los SNPs heterocigotos, Figura 9).

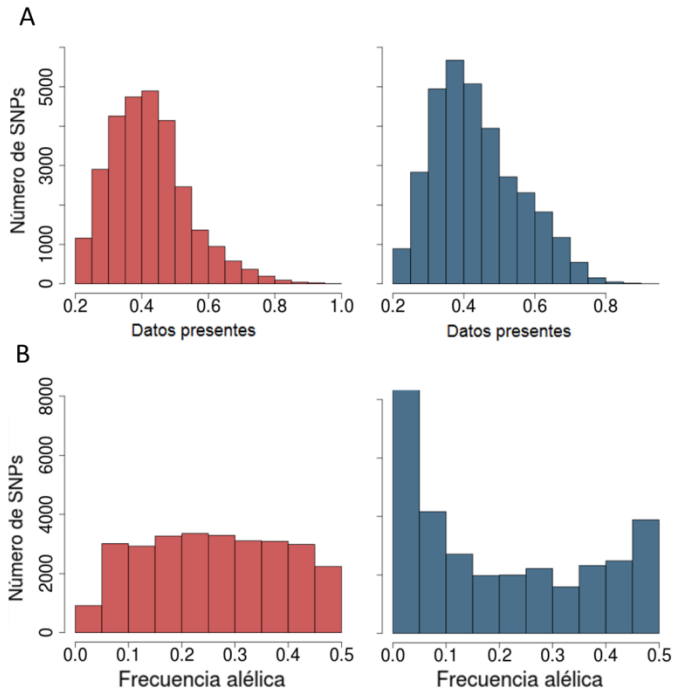



Figura 8: A) Número de SNPs según la proporción de individuos en los que esta presentes. B) Distribución de los alelos con menor frecuencia. En color rojo se presenta la distribución para los SNPs obtenidos por PipT y en azul para PipRef.

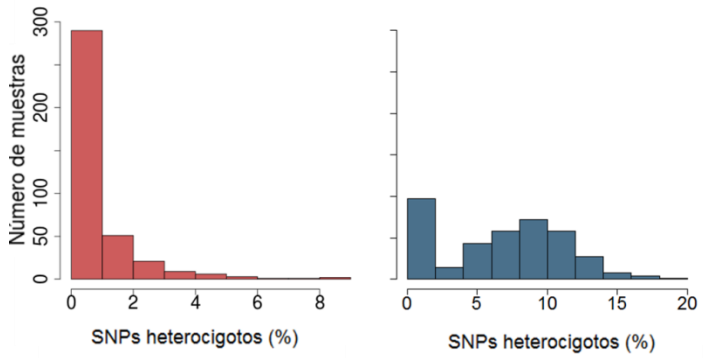



Figura 9: Proporción de SNPs heterocigotos en los 382 genotipos secuenciados. En color rojo se muestran los heterocigotos para PipT y en azul, para PipRef.

### 3.3. VALIDACIÓN DE LOS SNPs IDENTIFICADOS

A continuación se describen los resultados obtenidos con los métodos de validación de SNPs propuestos.

### 3.3.1. Análisis de correlación entre matrices genéticas

Se presentan los resultados obtenidos de las correlaciones entre las matrices de distancia genéticas calculadas con los subconjuntos de SNPs. Las matrices se obtienen con y sin imputación de los datos, por lo que se calcularon dos valores de correlaciones para cada pipeline.

En el estudio de la correlación entre las matrices de distancia genética, calculadas con los SNPs obtenidos por PipT (14.099 SNPs en cada una de las matrices), se obtuvo un valor de correlación de Pearson de 0,93 ($p<0,001$). Este resultado se obtuvo a través del cálculo de la matriz de parentesco con un elevado número de datos perdidos (aproximadamente 50 % de los datos). Por lo tanto, al imputar dichos datos (utilizando la media entre columnas y filas de la matriz), se observó una mejora en la correlación entre ambas matrices, obteniéndose un valor de correlación de 0,97 ( test de mantel $p<0,001$; $r^2 = 0,94$). Estos mismos análisis se realizaron con los datos de PipRef, y en el mejor de los casos (con imputación de datos), se obtuvo un valor $r^2$ de 0,82 ($r= 0.91$, $p<0.001$, Figura 10B).

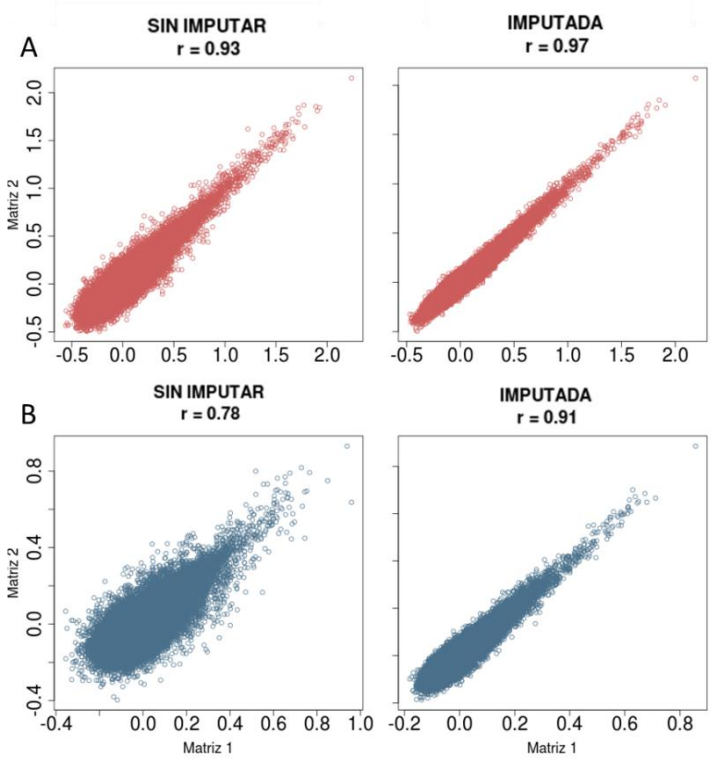



Figura 10: A) Correlación de las matrices genéticas sin (izq.) y con imputación (der.), para PipT. B) Correlación de las matrices genéticas sin (izq.) y con imputación (der.), para PipRef.

A continuación se analizó la correlación de las matrices de relaciones genéticas calculadas con el conjunto de los SNPs identificados en cada uno de los pipelines. La correlación de la distancia genética entre individuos en ambas matrices fue aceptable (r = 0.93, p<0.001), aunque presentó diferencias entre ambos conjuntos de datos. También se observaron diferencias en el rango de valores de distancia entre matrices; mientras que para PipRef los valores oscilaron entre -0,2 y 0,8, para PipT, se obtuvieron valores entre -0,5 y 2,0 (Figura 11).

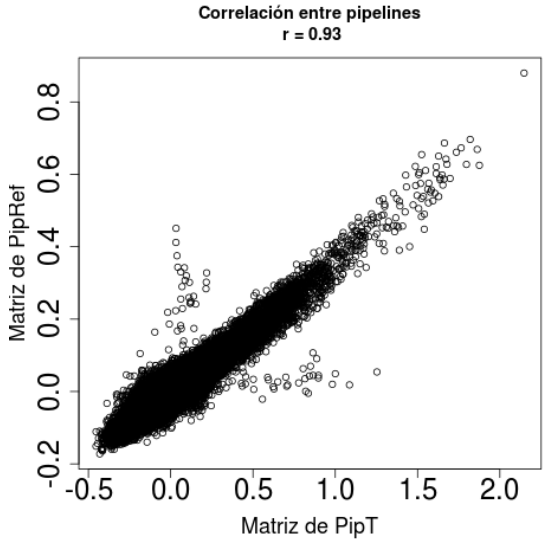



Figura 11. Plot de correlación entre las matrices de parentesco, calculadas utilizando la matriz de SNPs de cada pipeline.

Al realizar la comparación de datos de los SNPs identificados utilizando ambos pipelines, se observaron mayores correlaciones entre matrices de distancia genética para el PipT, el cual a su vez, presentó menores niveles de heterocigocidad en los SNPs identificados y una mejor distribución de las frecuencias alélicas. Por lo tanto, se continuó la validación de SNPs únicamente para la metodología PipT.

### 3.3.2. Análisis de la estructura poblacional

Se conoce tanto la información de procedencia como el pedigrí de los genotipos considerados en el estudio. Con el objetivo de relacionar estos datos con los datos genéticos y evaluar como se agrupan las muestras, se realizó un análisis de PCoA de la matriz de distancia genéticas calculada con el conjunto de los SNPs. A continuación, se calculó la matriz de

coancestría para los 177 individuos (que presentaban suficiente información de sus ancestros) del total de 182 materiales uruguayos genotipados y se correlacionó con la matriz genética.

En el análisis de coordenadas principales se identificaron 2 grupos, el grupo 1 formado principalmente por materiales uruguayos y el grupo 2 por materiales chilenos, si bien existe una proporción de materiales que se entrecruzan (Figura 12). La correlación entre la matriz de coancestría y la matriz de relaciones genéticas fue 0,51 ($p < 0.0001$). Los valores de la matriz de distancia genética calculada con los marcadores moleculares oscilaron entre 0,577 y 0,990, con media 0,682 y mediana 0,975, sin tomar en cuenta la diagonal. Por otro lado, la matriz de coancestría presentó valores entre 0,0003 y 0,779, con media 0,031 y mediana 0,061. El promedio del número de generaciones utilizadas para la construcción de la matriz COP fue de 18 generaciones, con un mínimo de 11 y un máximo de 26 generaciones.

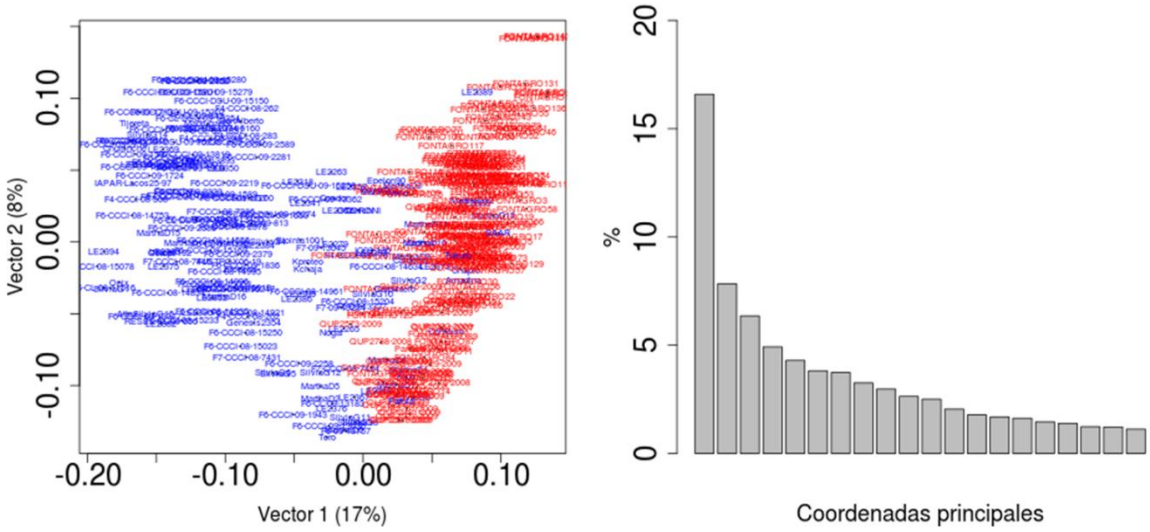



Figura 12. Análisis de coordenadas principales. En color azul se destacan las variedades de Uruguay y en rojo las de Chile. A la izquierda se presenta la proporción de varianza explicada por las primeras 20 coordenadas principales.

### 3.3.3. Mapeo de SNPs y análisis de desequilibrio de ligamiento (LD)

Se realizo un mapeo de los SNPs en las bases de datos disponibles (Poland et al. 2012a y base de datos del Institut National de la Recherche Agronomique (INRA)). En el mapeo de los SNPs en referencia a la base de datos de Poland et al. (2012a) se localizaron el 45 % de los

SNPs identificados (12.587); los SNPs se asignaron a un cromosoma y se ordenaron según los datos del mapa de bin de la base de datos. El alineamiento con la base de datos del URGI, INRA (http://urgi.versailles.inra.fr/Species/Wheat) nos permitió identificar el cromosoma en el cual se encuentran el 64 % de los SNPs (17.754) aunque no fue posible identificar la posición exacta, ni el orden de los SNPs dentro del cromosoma. Se identifican SNPs en todos los cromosomas y la distribución fue similar en ambas base de datos (Figura 13). Para los siete cromosomas se identificaron menos SNPs en el genoma D (Figura 13).

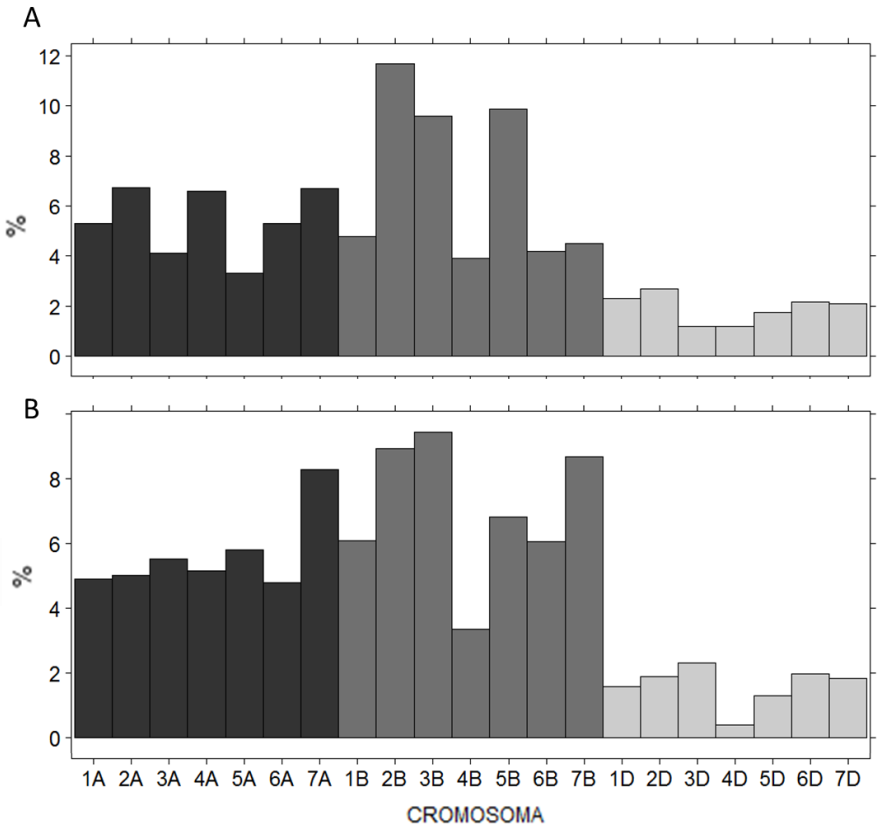



Figura 13. Distribución de los SNPs en los distintos cromosomas. A) Mapeo de SNPs en la base de datos de www.urgi.versailles.inra.fr/Species/Wheat/. B) Mapeo de SNPs en la base de datos Poland et al. (2012a).

En el análisis de LD para los SNPs mapeados se obtuvieron distintos grupos de ligamiento (Figura 14). En la Figura 15 se presenta el mismo análisis pero para la base de datos de URGI, la cual no permite ordenar los SNPs dentro del cromosoma. Por lo tanto, los grupos de ligamiento se distinguen con menor claridad y únicamente es posible identificar algún grupo aislado cuando se grafican conjuntamente varios cromosomas.

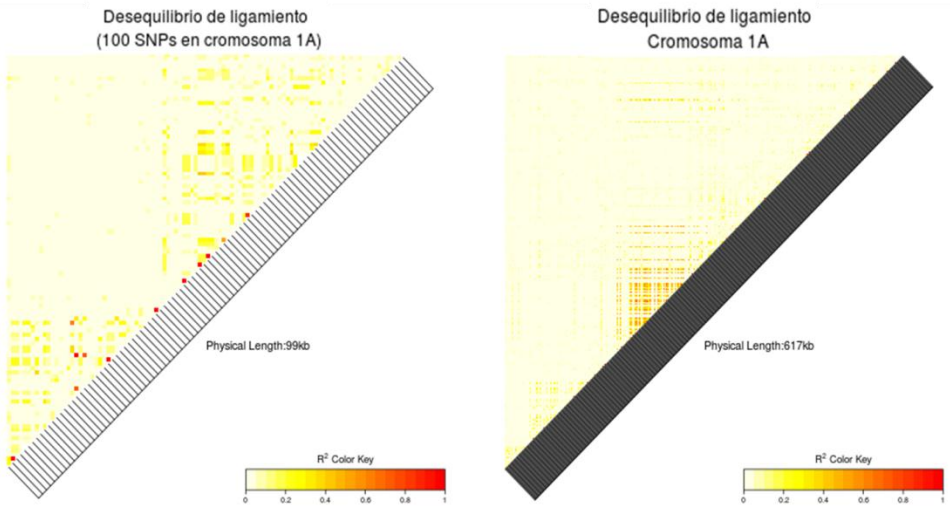



Figura 14. Análisis de desequilibrio de ligamiento. A) LD para los primeros 100 SNPs. B) LD a lo largo del cromosoma 1A (618 SNPs). El valor de $r^2$ se codifica con un código de colores que oscila de blanco-amarillo ($r^2$=0) a rojo ($r^2$=1).

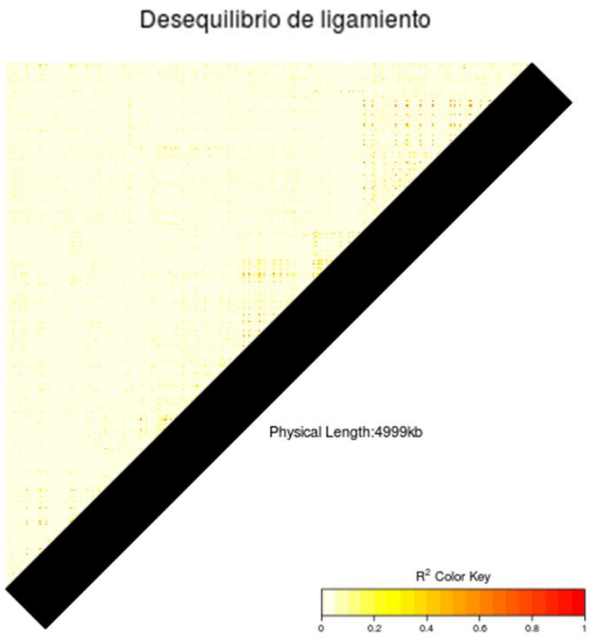



Figura 15. LD para los primeros 5000 SNPs que se mapearon utilizando el BLAST de URGI. La escala de colores que oscila entre blanco y el rojo indica el valor de $r^2$ de 0 a 1, respectivamente. Los 5000 SNPs comprenden los cromosomas 1A, 1B, 1D, 2A, 2B ,2D, 3A, 3B, 3D, 4A, 4B y 4D.

### 3.3.4. Predicciones fenotípicas

Las variables elegidas para realizar las predicciones fueron rendimiento y peso de mil granos, que son variables con distintas características de heredabilidad. Se estudiaron las variables individualmente en cada uno de los ambientes (riego y secano). Ambas características se ajustaron a una distribución normal (Figura 16).

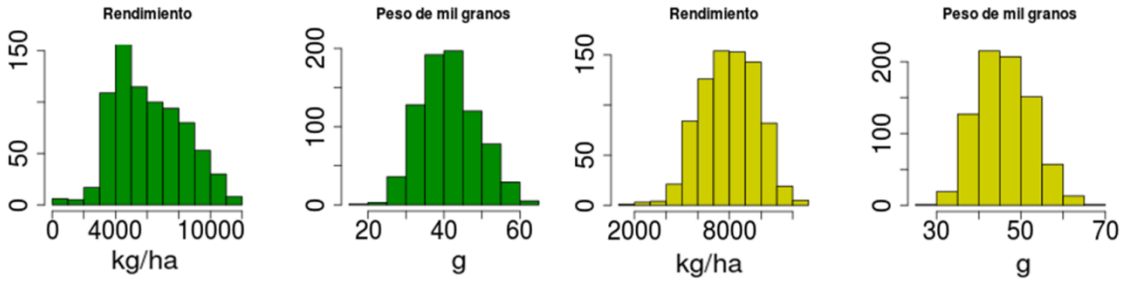



Figura 16. Distribución de las variables rendimiento y peso de mil granos. En verde se muestra Riego y en amarillo Secano.

El análisis de los residuales para los datos fenotípicos obtenidos en el campo (modelo de látice simple, LS), permitió observar la existencia de correlación entre ellos, lo cual es consistente con la presencia de variación espacial en el campo (Figura 17). Por lo tanto, se plantearon 3 modelos de corrección de dicha variación espacial: FC, BCA_MVNG y MVNG, con el objetivo de obtener las mejores medias ajustadas, que eliminen la correlación existente debido a la variación espacial y permitan obtener mejores predicciones.

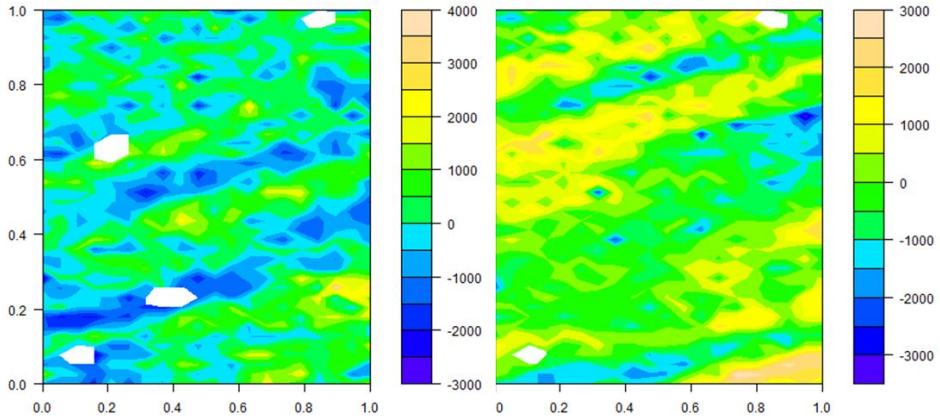


Figura17. Distribución espacial de los residuales ajustados con el diseño de látice simple (LS). En ambos ambientes, se asigna una coordenada espacial a cada valor de los residuos. Los colores indican los valores de los residuales: en azul valores negativos y en amarillo

valores positivos y en verde valores intermedios. A la izquierda se observan los valores obtenidos para los genotipos en condición de riego y a la derecha de secano.

En el análisis de las medias ajustadas por tratamiento, para los 4 modelos y las medias sin ajustar, se observó una menor desviación para BCA_MVNG y MVNG en el conjunto de los datos (Figura 18).

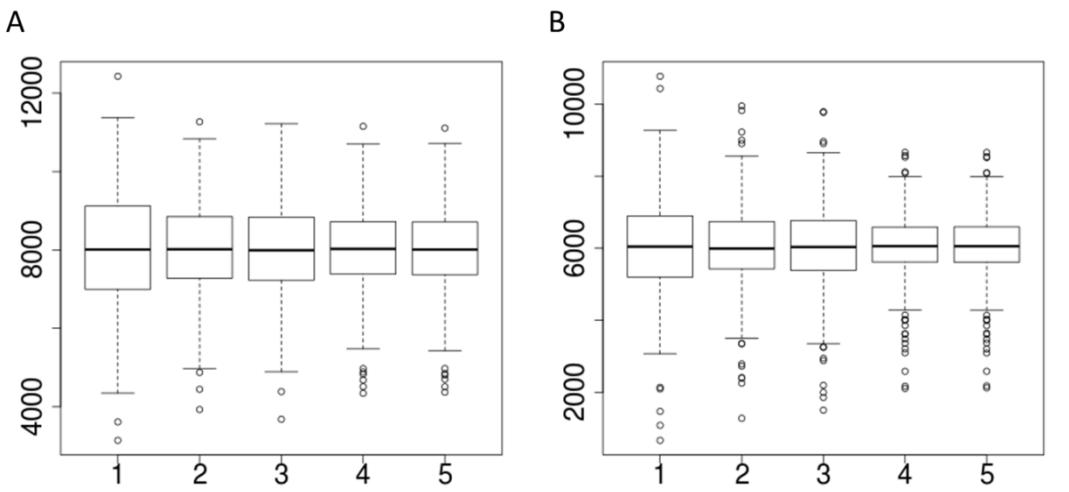



Figura 18. Medias ajustadas para los tratamientos sin ajustar-control (1) y para cada uno de los modelos evaluados: BI (2), FC (3), BCA_MNVG (4) y MVNG (5), en condiciones de riego (A) y secano (B).

Con el objetivo de evaluar la efectividad del ajuste de los distintos modelos, se graficó la distribución espacial de los residuales para cada modelo (Figura. 19, se grafica para riego a modo de ejemplo dado que secano presento similar comportamiento). Se observó que los modelos BCA_MVNG y MVNG corrigieron la tendencia encontrada en los residuales del modelo LS. El perfil de los residuales del modelo con corrección por filas y columnas fue similar al encontrado en el ajuste con el modelo del diseño experimental.

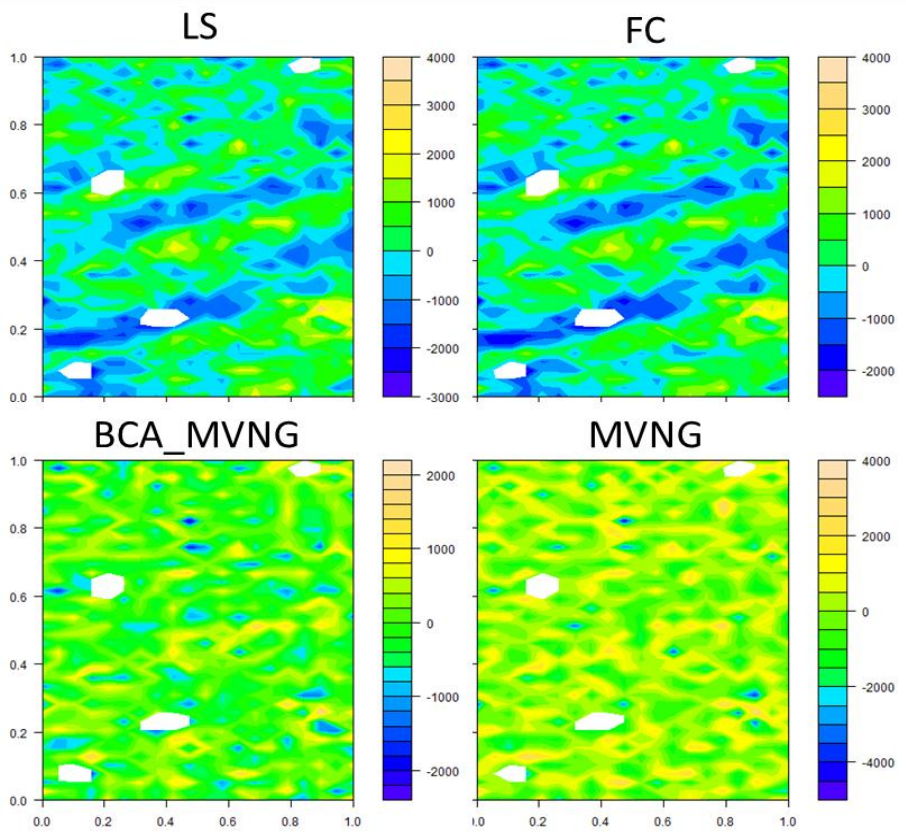



Figura 19. Distribución de los residuales obtenidos con cada uno de los modelos de ajuste. A cada valor de residual se le asigna la posición en el campo y con los colores se indica la magnitud.

Con el objetivo de comparar los modelos para cada una de las variables, se calculó la heredabilidad, obteniéndose un valor en 3 de los 4 modelos, dado que para el modelo MVNG no se conoce el dato de la varianza entre los tratamientos. La heredabilidad fue más alta para la variable peso de mil granos que para rendimiento, aumentando la misma en ambos casos al ajustar los datos fenotípicos con el modelo BCA_MVNG. Este aumento fue más significativo en el caso del rendimiento (Tabla 3). Luego de ajustar los modelos, se obtuvieron los BLUP (los mejores predictores lineales insesgados), a partir de los cuales se realizaron las predicciones en ambas variables (Tabla 4). La validación cruzada permitió evaluar la precisión de las predicciones, mediante correlaciones entre los GEBV y los BLUP obtenidos con los distintos modelos. Las predicciones presentaron mejor correlación para la variable peso de mil granos que para rendimiento. En cuanto a las predicciones de los distintos modelos, la mayor precisión fue la de los modelos que incluyen las medias móviles como

covariable, siendo MVNG el modelo que presentó los valores más altos de correlación (Tabla 4). En los modelos FC y LS se observan valores de correlación similares.

Tabla 3. Heredabilidad de las variables rendimiento y peso de mil granos, bajo condiciones de riego y secano.

| | **Rendimiento** | | **Peso de mil granos** | |
|---|---|---|---|---|
| **Modelos** | **H riego** | **H secano** | **H riego** | **H secano** |
| **LS** | 0,42 | 0,33 | 0,88 | 0,79 |
| **FC** | 0,44 | 0,37 | 0,88 | 0,81 |
| **BCA_MVNG** | 0,57 | 0,56 | 0,89 | 0,82 |

**FC**: ajuste con modelo de filas y columnas; **LS**: ajuste con el modelo de diseño un alfa látice simple; **BCA_MVNG**: ajuste con modelo de diseño de bloque completos al azar con medias móviles como covariable

Tabla 4. Correlaciones entre los GEBV y los BLUP en la población de validación, para rendimiento (Rend.) y peso de mil granos (P1000g), con los 4 modelos planteados y para las medias sin ajustar, en los ambientes evaluados (riego y secano).

| | | **Rend. riego** | **P1000g riego** | **Rend. secano** | **P1000g secano** |
|---|---|---|---|---|---|
| **Sin ajustar** | RR | 0,174±0,057 | 0,794±0,022 | 0,129±0,051 | 0,752±0,023 |
| | GAUSS | 0,337±0,090 | 0,794±0,022 | 0,285±0,108 | 0,754±0,025 |
| **FC** | RR | 0,221±0,061 | 0,755±0,021 | 0,171±0,068 | 0,766±0,019 |
| | GAUSS | 0,341±0,084 | 0,753±0,022 | 0,315±0,092 | 0,767±0,019 |
| **LS** | RR | 0,245±0,060 | 0,751±0,023 | 0,191±0,061 | 0,745±0,026 |
| | GAUSS | 0,361±0,074 | 0,751±0,031 | 0,324±0,078 | 0,748±0,027 |
| **BCA_MVNG** | RR | 0,246±0,077 | 0,752±0,022 | 0,219±0,060 | 0,764±0,019 |
| | GAUSS | 0,374±0,078 | 0,753±0,023 | 0,339±0,086 | 0,764±0,021 |
| **MVNG** | RR | 0,278±0,068 | 0,787±0,019 | 0,216±0,065 | 0,799±0,021 |
| | GAUSS | 0,396±0,070 | 0,787±0,020 | 0,357±0,084 | 0,799±0,02 |

Sin ajustar: medias fenotípicas con promedios simples; **FC**: ajuste con modelo de filas y columnas; **LS**: ajuste con el modelo de diseño un alfa látice simple; **BCA_MVNG**: ajuste con modelo de diseño de bloque completos al azar con medias móviles como covariable; **MVNG**: ajuste de un modelo de regresión lineal con medias móviles como variable regresora.

## 4. DISCUSIÓN

La mejora genética ha permitido incrementar los niveles de producción mundial de los cultivos a lo largo de la historia (Prakash, 2001). Los descubrimientos en el área genética y biotecnológica logrados durante el último siglo no solo aportaron las bases teóricas para comprender los mecanismos que operan en la mejora de los cultivos, sino que también brindaron nuevas herramientas para la manipulación de esos mecanismos. La primera herramienta biotecnológica utilizada fue la irradiación y modificación química de cultivos con el objetivo de crear variabilidad genética (Prakash, 2001). Hoy en día este objetivo se busca también mediante la transgénesis, que es un método más específico y permite mayor variación genética por expresión de genes provenientes de diferentes especies (Pakash, 2001; Abdul et al., 2009). Junto a las metodologías de transgénesis se aplican métodos de identificación y selección de genes por marcadores moleculares que son menos costosos y controversiales que los primeros (Zamir, 2001). La generación de nuevos cultivares depende directamente de la capacidad de mejorar la expresión de características complejas, es decir, influenciadas por varios genes (Khush, 2001). Sin embargo, los métodos hasta ahora existentes, que permiten asociar la información genética a la expresión fenotípica de estos caracteres presentan ciertas limitaciones, debido a las interacciones genéticas y al alto número de genes involucrados en la expresión de las variables (Verhoeven et al., 2010; Lu et al., 2011).

Por otro lado, el desarrollo tecnológico en el área de la secuenciación ha dado el puntapié inicial para la puesta a punto de técnicas de genotipado masivo, en las que se propone disminuir costos y simplificar la metodología de obtención de marcadores moleculares.

En el último tiempo se han llevado a cabo distintos trabajos, con el objetivo de obtener marcadores moleculares, utilizando dichas tecnologías en distintas especies, como ser *Aegilops tauschii* (You et al., 2011), soja (Deschamps et al., 2010; Wu et al., 2010; Varala et al. 2011), trigo (Trebbi et al., 2011) y frijol (Hyten et al., 2010c). Es importante destacar que la aplicación de esta metodología para la identificación de marcadores en el genoma de trigo presenta dificultades debido a las características complejas de su genoma. La presencia de cromosomas homeólogos y la alta proporción de regiones repetidas en genomas poliploides como el de trigo, son 2 obstáculos en la identificación de SNPs (Akhunov et al., 2009). En este sentido, para evitar la identificación de falsos SNPs por homeólogos o parálogos, es importante tener en cuenta que la diversidad entre estas regiones es mayor que la diversidad entre distintos cultivares de trigo. Se ha determinado que la frecuencia de encontrar un SNP

entre variedades de trigo oscila entre 1 en 200 a 613 pb, mientras que para regiones homeólogas, la misma asciende a 1 en 24 a 61 pb (Somers et al., 2003; Schnurbusch et al., 2007; Barker y Edwards, 2009). Recientemente se propuso el genotipado por secuenciación como una herramienta alternativa para la identificación de SNPs en trigo (Poland, et al. 2012a), sacrificando la "representación" del genoma en lo fragmentos re-secuenciados, por una mayor cobertura de cada uno de los fragmentos (menos fragmentos que se secuencien más veces). Otras alternativas en este sentido se plantean con la construcción de librerías de representación reducida del genoma, las cuales permiten disminuir aún más la "representación" de los fragmentos aumentando su cobertura (Hyten et al., 2010b; Varala et al., 2011).

La necesidad de desarrollar nuevas metodologías para la selección de características complejas junto a los avances en el área del genotipado, han llevado al desarrollo de nuevas metodologías de selección, como ser la selección genómica. Cabe destacar que en estas técnicas no solo es importante la calidad de los datos genotípicos, sino también la de los datos fenotípicos utilizados para entrenar los modelos de selección (Cabrera-Bosquet et al., 2012).

Con el objetivo de incorporar estas nuevas metodologías a los programas de mejoramiento en Uruguay, es que se propuso aplicar GBS en 384 genotipos de trigo ajustando un método bioinformático de identificación de SNPs para ser utilizando en SG. El análisis de los datos fenotípicos obtenidos a campo permitió introducir en el trabajo la evaluación de nuevos modelos de corrección espacial, aportando una herramienta para su aplicación en ensayos de evaluación de genotipos en programas de mejoramiento genético.

### 4.1. IDENTIFICACIÓN DE SNPs

Para el análisis de la información brindada por el secuenciador es recomendable estudiar la distribución de calidad de la bases en el conjunto de fragmentos obtenidos; lo que permite detectar la existencia de algún problema durante la secuenciación (DeWit et al., 2012). En este trabajo, la calidad de los fragmentos es elevada hasta el final de la secuenciación, disminuyendo hacia el extremo 3'. A su vez, la media y el primer cuartil, de la distribución de la calidad de la bases en el conjunto de fragmentos, se mantuvieron cercanas a 30, que indican una probabilidad de error de uno en mil (DeWit et al, 2012).

El análisis de profundidad de las secuencias (número de veces que se encuentra repetida) por línea del secuenciador indica que, aproximadamente un 2,5 % del total de secuencias se encuentran repetidas. Cada una de las muestras está representada por el 1 % del total (millón y medio de secuencias) de fragmentos de ADN en la línea del secuenciador. Lo anterior indica que hay una gran proporción de información que se descarta, dado que secuencias de única repetición no son tomadas en cuenta en los análisis posteriores, debido a que existe una alta probabilidad de que se deban a errores durante la secuenciación o amplificación (Elshire etal., 2011; Varala et al., 2011). Por lo tanto el incrementar la precisión durante la secuenciación y evitar los pasos de amplificación son dos factores a tener en cuenta para un uso más eficiente de la información obtenida (Metzker, 2010; Niedringhaus et al., 2011).

La identificación de SNPs utilizando las metodologías PipRef y PipT produjo resultados que se diferencian manteniendo algunos puntos en común. La principal diferencia entre ambas radica en el método de alineación de los fragmentos durante el proceso de análisis de posibles variaciones. En PipRef fue necesario contar un ensamblado de secuencias para utilizarlo como genoma de referencia, ya que en el caso del trigo aún no se dispone de esta herramienta. Es importante destacar que la calidad de este ensamblado es una limitante en la aplicación de esta metodología. La limitaciones en este caso están dadas por la cobertura del genoma que está representada en el ensamblado y por lo posibles errores existentes en el mismo (Hyten et al., 2010c; Alkan et al., 2011; Helyar et al., 2012). En el caso de PipT no es necesario contar con un genoma de referencia, sino que se realizan alineamientos de a pares entre las secuencias generadas. Una limitante de esta metodología es que se requiere un buen número de secuencias repetidas para asegurar el SNP (Poland et al.2012a y 2012b), dado que ambos alelos están determinados únicamente por el conjunto de secuencias re-secuenciadas (Ratan et al., 2010; Elshire et al., 2011). En este caso el contar con un genoma de referencia permitiría ser menos exigente en este punto. En el caso de PipRef es necesario contar con al menos 2 secuencias para llamar a cada genotipo (asignarle un alelo u otro a la muestra), tomando también en cuenta SNPs en función de la profundidad media en el conjunto de individuos entre un máximo y un mínimo. El valor medio mínimo de profundidad permite asegurarse que la lecturas se encuentran distribuidas en cierta proporción de las muestras y que no se encunetran concentradas en pocos individuos (Trebbi et al., 2011). En este sentido se establece un umbral de valor medio máximo de profundidad, a modo de eliminar secuencias que provengan de zonas repetidas del genoma (Wu et al., 2010). Estos son parámetros que se pueden modificar en busca de optimizar la identificación de SNPs (Nielsen

et al., 2011). Otra de la diferencias importantes entre ambas metodologías es el tamaño de fragmento, para PipT es de 64pb y en el caso de PipRef un promedio de 80–85 pb. Cuanto más largo es el fragmento, con mayor exactitud se produce el alineamiento entre secuencias, asignando la dirección correcta en el caso de alinear contra una referencia (Harismendy et al., 2009). Los SNPs identificados mediantes ambos pipelines mostraron un perfil de proporción de SNPs por muestra similar. Una característica negativa de GBS es que se genera una elevada cantidad de datos faltantes (Elshire et al., 2011; Poland et al., 2012a y 2012b); la mayoría de la muestras presentaron un promedio del 50 % de los SNPs sin datos, coincidiendo con los resultados obtenidos por Poland et al. (2012a y 2012c). Una forma de resolver esta dificultad es genotipar un gran número de muestras que permitan inferir información unas a otras mediante la imputación de los datos. Poland et al. (2012c) describen distintas metodologías de imputación de datos, midiendo la eficiencia de la imputación simulando datos perdidos.

Los SNPs identificados por ambos pipelines presentaron distribuciones similares de datos faltantes en el conjunto de muestras. La principal diferencia entre ambas metodologías radica en las frecuencias alélicas de los SNPs y el número de SNPs heterocigotos. Al establecer un filtro por máxima frecuencia alélica 0,01 se filtraron alelos presentes en menos del 1% de las muestras (Kim et al., 2011). La frecuencia alélica en el conjunto de SNPs identificados por PipT presentó una distribución homogénea, con un menor número de SNPs de frecuencia alélica entre 0,01 y 0,05. Por otro lado, con PipRef se obtuvo una alta proporción de SNPs con valores de frecuencia alélica entre 0,01 y 0,05. La proporción de SNPs heterocigotos también aumenta significativamente al utilizar PipRef, a pesar de que en trigo (especie autógama) se espera baja proporción de heterocigosis (Enjalbert y David, 2000).

### 4.2. VALIDACIÓN DE SNPs

Para la validación de los SNPs se utilizaron distintas estrategias para decidir si la información de los marcadores obtenida a partir del conjunto de datos, representa una información coherente acerca de los individuos cuyas muestras han sido secuenciadas. Todas las metodologías propuestas utilizan el conjunto de los datos para su validación, ya sea mediante el cálculo de matrices de distancia genética, mediante mapeo y análisis de desequilibrio de ligamiento de los SNPs o a través de la realización de predicciones fenotípicas. En diferentes

trabajos es común realizar validación mediante la amplificación y re-secuenciación por el método de Sanger de un determinado número de SNPs escogidos al azar, en distintos genotipos (Wu eta;l., 2010; Deschamps et al., 2010; You et al., 2011). En este trabajo se descartó la aplicación de este método de validación, ya que al contar con 28.199 SNPs detectados sería necesario validar al menos de 100 SNPs en por lo menos 8 genotipos. Lo anterior será costoso y aportaría la misma información que los métodos de validación planteados en este trabajo.

### 4.2.1. Correlación entre matrices genéticas

Las distancias genéticas entre los individuos, calculadas a partir de subgrupos de SNPs divididos aleatoriamente, mostraron buena correlación, que aumentó al imputar los datos faltantes. El valor de correlación de las matrices fue mayor para los SNPs identificados mediante PipT que mediante PipRef. Lo anterior es un indicador de que los SNPs identificados en PipRef estarían introduciendo errores en el cálculo de distancias genéticas, probablemente debido a un mayor número de falsos SNPs (Hamblin et al., 2007; Chao et al., 2009). Sin embargo, las distancias genéticas calculadas con los SNPs identificados mediante PipT correlacionó muy bien ($r=0.93$, $p<0.001$) con las distancias calculadas con los SNPs identificados en PipRef. Esto nos indica que ambas matrices poseen información común (Chao et al., 2009). En base a los descrito anteriormente y considerando también un mayor número de heterocigotos y alelos con bajas frecuencias en el caso de PipRef, se decidió seleccionar la metodología de PipT para continuar el proceso de validación. Sin embargo, cabe considerar que la limitante de la ausencia del genoma de referencia en el caso de trigo puede ser levantada conforme surjan nuevas investigaciones en la temática. PipRef ha sido aplicado exitosamente al cultivo de soja (con algunas variante en los filtros por tratarse de otra especie) que cuenta con un genoma de referencia (com. pers., F. Belzile, Universidad de Laval Quebec).

### 4.2.2. Correlación entre SNPs y Pedigrí

El análisis de coordenadas principales, a partir de la matriz de distancias genéticas entre individuos, si bien mostró separación entre los materiales chilenos y uruguayos, ambos

grupos se mezclan. La mezcla de materiales entre ambos grupos se explica por la presencia de ancestros comunes a ambos programas de mejoramiento, provenientes de CIMMYT. Lo anterior es una característica deseable para su aplicación en genética asociativa o selección genómica, evitando los problemas que generan una población altamente estructurada (Jannink y Walsh, 2002; Breseghello y Sorrells, 2006).

La correlación entre la matriz de coancestría calculada con el pedigrí y la matriz de parentesco genética es mayor que la descrita por Almanza-Pizón et al. (2003) para marcadores SSR y AFLP, de todos modos siguen siendo valores de correlación medios ($r=0.51$, $p<0.001$). Esto podría deberse a las limitaciones para establecer relaciones entre líneas por su pedigrí, dada la falta de información para extender los pedigrís de cada genotipo. En general en trigo se ha visto que los marcadores moleculares son más sensibles para establecer relaciones entre individuos (Manifesto et al., 2001; Almaza-Pinzón et al., 2003; Fufa et al., 2005). Los valores de distancia genéticas de las matrices de coancestría son mayores (más cercanos a 0, valor que representa una distancia genética infinita) que los valores de las matrices de parentesco genéticas. Si bien estas limitaciones no permite llegar a una conclusión definitiva, la correlación entre marcadores y pedigrí es mayor a la hallada con marcadores SRR y AFLP (Manifesto et al., 2001; Almaza-Pinzón et al., 2003).

Por otro lado, la correlación entre matrices de parentesco genéticas y coancestría depende mucho de la calidad de los datos con los que se trabaje. Se supone que si tenemos una cobertura ideal de marcadores y una base de datos de pedigrí completa, estas dos matrices tenderían a correlacionar muy bien, pero estas situaciones son ideales. Crossa et al. (2010) muestra que el uso de las dos matrices aportan información complementaria en los análisis que incluyen datos genéticos.

### 4.2.3. Mapeo de SNPs

El mapeo de SNPs en una base de datos es una herramienta útil en la validación de estos marcadores. Referenciar una alta proporción de los SNPs a un cromosoma permite tener una idea de cómo es su distribución en los distintos cromosomas y genomas (Chao et al., 2009). A su vez la comparación del nivel de polimorfismo identificado en los cromosomas con el hallado en otros trabajos nos podría aportar nueva evidencia en la validación. Por otro lado, el orden de los marcadores dentro del cromosoma y el cálculo del valor de $r^2$ entre los

marcadores permitieron identificar la formación de grupos de ligamiento dentro del cromosoma. Estas asociaciones aportan evidencias que ayudan confirman la veracidad de los SNPs identificados (Chao et al., 2010)

El genoma D presentó significativamente menor número de SNPs que los otros dos homeólogos A y B, confirmando lo descrito en otros estudios llevados a cabo en trigo (Chao et al., 2009 y 2010; Allen et al., 2011). A su vez, en la gran mayoría de los cromosomas el genoma B presenta mayor cantidad de SNPs que el genoma A, situación que también se ha descrito en trabajos previos (Gu et al. 2004; Chao et al., 2009).

Una limitación en este análisis es la imposibilidad de establecer una dirección genética de los SNPs en los cromosomas, debido a que las bases de datos utilizadas para mapeo no poseen dicha información. Sin embargo, al ordenar los SNPs según el mapa de bins de Poland et al. (2012a), se pudo identificar la formación de grupos de ligamento a lo largo del cromosoma. Lo anterior es un indicio de la identificación verdaderas variaciones entre individuos, dado que se espera que los falsos SNPs sean espurios en el conjunto de individuos (Gupta et al., 2005).

### 4.2.4. Predicciones fenotípicas

Como último método de validación se planteó realizar predicciones fenotípicas y estudiar su precisión mediante validación cruzada. Para realizar las predicciones fenotípicas a partir de los 28.199 SNPs identificados, se utilizaron los datos obtenidos a campo de 381 del total de las variedades genotipadas. Las variables escogidas en la validación fueron rendimiento y peso de mil granos. El rendimiento, es una variable de interés y de difícil predicción, dado que está influenciada por muchos genes que actúan sobre diversas características (Wu et al., 2012). El peso de mil granos es uno de los componentes de rendimiento, presentó un mayor valor de heredabilidad que el anterior, lo que presupone mejores predicciones (Iwata y Jannink, 2011).

A diferencia de lo encontrado en las simulaciones planteadas por Gezan et al. 2010 el diseño de filas y columnas no mostró mejoras con respeto al diseño de látice simple realizado a campo, no siendo adecuado para la corrección de variación espacial detectada en los datos. Coincidiendo con los trabajos de Townley-Smith y Hurd (1973) y Gezan et al. (2010) se

propuso incluir en los modelos las medias móviles como covariables, seleccionando el modelo lineal que incluye dicha covariable como único componente regresor.

Un diseño a campo en el cual se controle el efecto diferencial, de los principales efectos ambientales, sobre los genotipos para evaluarlos aplicando la corrección según el diseño planteado; es el escenario ideal para llevar a cabo los ensayos (Edmondson, 2005). Sin embargo, por diversos factores que van desde el no conocimiento del campo hasta limitaciones en la disponibilidad de materiales, hacen que el planteo de un buen diseño previo no siempre sea posible (Cullis et al., 2006; Leiser et al., 2012). En los últimos tiempos se han publicado algunos trabajos que proponen distintos modelos para ajustar la variación espacial post obtención de los datos. Los principales modelos de corrección espacial existentes son el ajuste por filas y columnas, asociando este dato a cada valor fenotípico (Piepho y Williams, 2010), ajuste mediante la correlación de errores con distintas estructuras de covarianza (Federer, 2003), ajuste aplicando el método de vecinos más cercanos (Papadakis, 1937; Müller et al., 2010) y ajuste usando medias móviles (Townley-Smith y Hurd, 1973; Peiris et al., 2008). Para seleccionar los modelos es importante considerar el tipo de variación espacial encontrada (local, general, con tendencia, en focos, etc), ya que los distintos métodos muestran mejores o peores performances dependiendo del tipo de variación existente (Gezan et al., 2010)

Utilizando un modelo de regresión lineal con medas móviles como covariable, lo valores de precisión en las predicciones de rendimiento fueron menores que los obtenidos por Crossa et al. (2010) para trigo y mayores que los presentados por Heffner et al. (2011) y Poland et al. (2012a). En cambio para la variable peso de mil granos las correlaciones fueron mayores a las encontradas en Poland et al. (2012c). Es importante destacar, que la precisión en las predicciones depende, entre otros factores, de la calidad de los datos fenotípicos; la relación entre la población de entrenamiento y validación: tipo de población (emparentada o diversa) y del número de individuos que componen cada población; de los modelos de predicción utilizados; de la especie en estudio y de la variable estudiada (Crossa et al., 2010; Heffner et al., 2011; Burgueño 2012; Nakaya e Isobe, 2012; Poland et al., 2012c; Resende et al., 2012).

Heffner et al. (2010) analizaron las condiciones de los programas de mejoramiento genético de maíz y trigo invernal de la Universidad de Cornell. Estos trabajos han demostrado que mediante la aplicación de selección genómica, obteniendo una precisión de 0.30 en la validación de las predicciones de rendimiento en trigo, es posible obtener mejores ganancias

genéticas que con la aplicación de la selección clásica por marcadores. Los resultados de este trabajo son un indicio promisorio para la futura aplicación de selección genómica, dado los valores de correlación obtenidos en las validaciones cruzadas (0.40 y 0.36 para riego y secano, respectivamente).

Una de las limitaciones que presentan las predicciones realizadas con los datos fenotípicos radica en que la información proviene de dos ambientes (riego y secano) con el mismo patrón de correlación entre los residuales. Aumentando la cantidad de ambientes y mediante un análisis de multi-ambiente las predicciones obtenidas podrían mejorar (Burgueño et al., 2012). Otra información que puede ser incluida en el modelo con el objetivo de mejorar las predicciones es la matriz de coancestría calculada con los datos de pedigrí, como se demuestra en Crossa et al., (2010).

Finalmente, cabe destacar que el modelo de predicción puede ser ajustado utilizando distintas metodologías que tomen en cuenta de diferencialmente el efecto de los marcadores. Alguno de estos métodos es capaz de asignar valores diferenciales dependiendo de la localización de los marcadores (Crossa et al. 2010). En nuestro caso los modelos se ajustaron con el método de G-BLUP o *ridge regresión* (RR), el cual asume que todos los marcadores tienen la misma distribución con media 0 y la misma varianza, y con el método de GAUSS. Este último, utiliza un *kernel* diferente para el cálculo de la matriz de Kinship, que permite considerar interacciones epistáticas entre marcadores (Gianola y van Kaam, 2008; Endelman; 2011)

La Selección Genómica es un método que propone acortar los tiempos y aumentar la eficiencia de mejoramiento, pero su aplicación depende de la calidad de los datos fenotípicos y genotípicos obtenidos. Este trabajo permitió validar un procedimiento de identificación de SNPs para su aplicación directa en un modelo de selección genómica. Para que esto sea exitoso, es condición necesaria aunque no suficiente, contar con un buen diseño a campo que permita obtener datos fenotípicos confiables. En este sentido, contar con herramientas que permitan ajustar y corregir datos fenotípicos post-diseño puede mejorar mucho la precisión de la selección.

## 5. CONCLUSIONES

La aplicación de genotipado por secuenciación como herramienta para la identificación de SNPs fue exitosa en variedades de trigo provenientes de un programa de mejoramiento, identificándose un alto número de polimorfismos de una sola base que están distribuidos en los distintos cromosomas, formando grupos de ligamiento.

El método de genotipado por secuenciación permitió obtener una densidad adecuada de marcadores genéticos, lo que lo convierte en una herramienta promisoria para su aplicación en un programa de mejoramiento genético de trigo. La mayor limitación en la técnica radica en la disponibilidad de datos en relación al genoma de trigo, lo cual limita su aplicación para mapeo asociativo o análisis de QTL en trigo.

El análisis de las variables mediante la corrección de la variación espacial existente, permitió potenciar las predicciones realizadas mediante selección genómica en el cultivo de trigo.

Si bien la Selección Genómica es una metodología que continúa en desarrollo, este trabajo aporta evidencias que indican que sería un método aplicable en los programas de mejoramiento genético de trigo.

## 6. BIBLIOGRAFÍA


Abdul, R., Zhiying, M. and Wang, H. (2009). Genetic Transformation of Wheat (Triticum aestivum L): A Review. Triticeae Genomics and Genetics. 1(2):1–7.

Ahn, S. J., Costa, J. and Emanuel, J. R. (1996). PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. Nucleic acids research. 24(13):2623–5.

Akbari, M., Wenzl, P., Caig, V., Carling, J., Xia, L., Yang, S., Uszynski, G., Mohler, V., Lehmensiek, A., Kuchel, H., et al. (2006). Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. TAG Theorical and Applied Genetics. 113(8):1409-1420.

Akey, J., Jin, L. and Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? European journal of human genetics: EJHG. 9(4):291–300.

Akhunov, E. D., Goodyear, A. W., Geng, S., Qi, L.-li, Echalier, B., Gill, B. S., Gustafson, J. P., et al. (2003). The Organization and Rate of Evolution of Wheat Genomes Are Correlated With Recombination Rates Along Chromosome Arms. Plant and Soil. 753-763.

Akhunov, E., Nicolet, C. and Dvorák, J. (2009). Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. TAG. Theorical and Applied Genetics. 119(3), 507-17.

Alkan, C., Sajjadian, S. and Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. Nature Methods. 8(1):61–65.

Allan, R.E. and Vogel O.A. (1964). F2 Monosomic Analysis of Coleoptile and First-Leaf Development in Two Series of Wheat Crosses. Crop Science. 4(3):338-339.

Allen, A. M., Barker, G. L. a, Berry, S. T., Coghill, J. a, Gwilliam, R., Kirby, S., Robinson, P., et al. (2011). Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). Plant biotechnology journal. 9(9):1086–99.

Almaza-Pinzón, M.I., Khairallah, M., Fox, P.N. and Warburton, M.L. (2003) Comparison of molecular markers and coefficients of parentage for the analysis of genetic diversity among spring bread wheat accessions. Euphytica. 130:77-86.

Andaya, V. C. and Mackill, D. J. (2003). Mapping of QTLs associated with cold tolerance during the vegetative stage in rice. Journal of experimental botany. 54(392):2579–85.

Ardlie, K.J.; Kruglyak, L. and Seielstad, M. 2002. Patterns of linkage disequilibrium in the human genome. Nature reviews. Genetics. 3:299-309.

Baldo, A.M., Yang, Y., Ramming, D.W. and Zhong G-Y. (2012) Genotyping by Sequencing in Grape - Stacks or Tassel?. International Plant and Animal Genome Conference.[En línea]. https://pag.confex.com/pag/xx/webprogram/Paper1731.html

Barker, G.L.A and Edwards, K.J. (2009). A genome-wide analysis of single nucleotide polymorphism diversity in the world's major cereal crops. Plant Biotechnology Journal. 7: 312-317.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., et al. (2010a). Galaxy: a web-based genome analysis tool for experimentalists. Current protocols in molecular biology, Chapter 19(January), Unit 19.10:1-21.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J. and Nekrutenko, A. (2010b). Manipulation of FASTQ data with Galaxy. Bioinformatics (Oxford, England). 26(14):1783-5.

Bnejdi, F. and Gazzah, M. E. (2010). Epistasis and genotype-by-environment interaction of grain protein content in durum wheat. Genetics and molecular biology. 33(1):125–130.

Bonfanti, D. (2010) Elites, Estado y técnicos extranjeros en los procesos de innovación agrícola en el Uruguay de los centenarios (1910-1930). Revista Encuentros Latinoamericanos. 10/11. Disponible en http://enclat.fhuce.edu.uy/index.php (verificado 05 Nov 2012)

Bonnett, D. G., Rebetzke, G. J. and Spielmeyer, W. (2005). Strategies for efficient implementation of molecular markers in wheat breeding. Molecular Breeding. 15(1):75–85.

Borlaug, N.E. (1983). Contribution of conventional plant breeding to food production. Science. 219(4585):689-693

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. a, Butler, T., Di Ventra, M., et al. (2008). The potential and challenges of nanopore sequencing. Nature biotechnology. 26(10):1146-53.

Breseghello, F. and Sorrells, M. E. (2006). Association Analysis as a Strategy for Improvement of Quantitative Traits in Plants. Crop Breedng and Genetics. 46:1323–1330.

Buerstymayr, H., Ban, T. and Anderson, J. A. (2009). QTL mapping and marker-assisted selection for Fusarium head blight resistance in wheat: a review. Plant Breeding. 128:1-26.

Burgueño, J., de los Campos, G., Weigel, K. and Crossa, J. (2012). Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. Crop Science. 52(2):707.

Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Dolors Serret, M. and Araus, J. L. (2012). High-throughput phenotyping and genomic selection: The frontiers of crop breeding converge. Journal of integrative plant biology. 00(0):1-9.

Carey, G. (2000). Quantitative Gnetics: II – Advanced Topics. Available: http://psych.colorado.edu/~carey/hgss/hgssadvanced/Chapter19Advanced.pdf

Castro, A.J., Gamba, F., German, S., Gonzales, S., Hayes, P.M., Pereyra, S. and Perez, C. (2012) Quantitative trait locus analysis of spot blotch and leaf rust resistance in the BCD47 × Baronesse barley mapping population. Plant Breeding. 131(2):258-266.

Chao, S., Zhang, W., Akhunov, E., Sherman, J., Ma, Y., Luo, M.-C. and Dubcovsky, J. (2009). Analysis of gene-derived SNP marker polymorphism in US wheat (Triticum aestivum L.) cultivars. Molecular Breeding. 23(1):23–33.

Chao, S., Dubcovsky, J., Dvorák, J., Luo, M.-C., Baenziger, S. P., Matnyazov, R., Clark, D. R., et al. (2010). Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (Triticum aestivum L.). BMC genomics. 11(1):727.

Chen, X., Min, D., Yasir, T. A. and Hu, Y.-G. (2012). Genetic Diversity, Population Structure and Linkage Disequilibrium in Elite Chinese Winter Wheat Investigated with SSR Markers. PloS one. 7(9):e44510

Chenu, K., Cooper, M., Hammer, G. L., Mathews, K. L., Dreccer, M. F. and Chapman, S. C. (2011). Environment characterization as an aid to wheat improvement: interpreting genotype-environment interactions by modelling water-deficit patterns in North-Eastern Australia. Journal of experimental botany. 62(6):1743–55.

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B. and 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.

Condón, F., Gustus, C., Rasmusson, D. C. and Smith, K. P. (2008). Effect of Advanced Cycle Breeding on Genetic Diversity in Barley Breeding Germplasm. Crop Science. 48(3):1027.

Cornide, M. T. (2001). LA GENETICA VEGETAL, EL MEJORAMIENTO Y LA SOCIEDAD. Cultivos Tropicales. 22(3):73–82.

Coster, A. (2011): Pedigree: Pedigree 2009. [R package version 1.2]

Crossa, J., de los Campos, G. D. L., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics. 186(2):713-24.

Cuesta-Marcos, A., Casas, A. M., Hayes, P. M., Gracia, M. P., Lasa, J. M., Ciudad, F., Codesal, P., et al. (2009). Yield QTL affected by heading date in Mediterranean grown barley. Plant Breeding. 128(1):46–53.

Cullis, B. R., Smith, a. B. and Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. Journal of Agricultural, Biological, and Environmental Statistics. 11(4):381–393.

Danecek, P., Auton, A., Abecasis, G., Albers, C. a, Banks, E., DePristo, M. a, Handsaker, R. E., et al. (2011). The variant call format and VCFtools. Bioinformatics (Oxford, England). 27(15):2156–8.

Davey, J. W., Hohenlohe, P. a, Etter, P. D., Boone, J. Q., Catchen, J. M. and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature reviews. Genetics. 12(7), 499-510.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., et al. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 182(1):375–85.

DePristo, M. a, Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. a, et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 43(5):491-8.

DeWit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., Therkildsen, N. O., et al. (2012). The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. Molecular ecology resources. 12(6):1058–67.

Deschamps, S., la Rota, M., Ratashak, J. P., Biddle, P., Thureen, D., Farmer, A., Luck, S., et al. (2010). Rapid Genome-wide Single Nucleotide Polymorphism Discovery in Soybean and Rice via Deep Resequencing of Reduced Representation Libraries with the Illumina Genome Analyzer. The Plant Genome Journal. 3(1):53.

Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. Nature, 418(6898):700–707

DIEA-MGAP. (2011) Anuario 2011. Disponible en http://www.mgap.gub.uy/portal/hgxpp001.aspx?7,5,27,O,S,0,MNU;E;2;16;10;6;MNU (verificado 21 Oct. 2012)

Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic acids research. 36(16):e105.

Dreisigacker, S., Zhang, P., Warburton, M. L., Skovmand, B., Hoisington, D. and Melchinger, A. E. (2005). Genetic Diversity among and within CIMMYT Wheat Landrace Accessions Investigated with SSRs and Implications for Plant Genetic Resources Management. Crop Science. 45(2):653.

Dubcovsky, J., Luo, M.-C., Zhong, G.-Y., Bransteiter, R., Desai, A., Kilian, A., Kleinhofs, A. and Dvorák, J. (1996). Genetic map of diploid wheat, Triticum monococcum L. and its comparison with maps of Hordeum vulgare L. Genetics. 143:983-999.

Dvorák, J. and Zhang, H. B. (1990). Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. Proceedings of the National Academy of Sciences of the United States of America. 87(24):9640-4.

Dvorák J, di Terlizzi P, Zhang HB, Resta P (1993) The evolution of polyploid wheats: identiWcation of the A genome donor species. Genome. 36:21–31.

Edmondson, R. N. (2005). Past developments and future opportunities in the design and analysis of crop experiments. The Journal of Agricultural Science. 143(1):27–33.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. a, Kawamoto, K., Buckler, E. S. and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one. 6(5):e19379.

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome Journal. 4(3):250.

Enjalbert, J. and David, J. L. (2000). Inferring recent outcrossing rates using multilocus individual heterozygosity: application to evolving wheat populations. Genetics. 156(4):1973–82.

Erlich, Y., Mitra, P. P., Mccombie, W. R. and Hannon, J. (2010). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nature Methods, 5(8), 679-682.

FAO-STAT. (2010). Prodaction. Disponible en http://faostat.fao.org/site/339/default.aspx (verificado el 21 Oct. 2012).

Federer, W. T. (2003). Exploratory Model Selection for Spatially Designed Experiments – Some Examples. Journal of Data Science. 1:231–248.

Flannery, K. V. (1982). Guilá Naquitz: Archaic Foraging and Early agriculture in Oaxaca, Mexico. Academic Press, New York.

Flint-García, S. A., Thornsberry, J. M. and Buckler, E. S. (2003). Structure of linkage disequilibrium in plants. Annual review of plant biology, 54, 357–74.

Flint-García, S. A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., Doebley, J., et al. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. The Plant journal: for cell and molecular biology. 44(6):1054–64.

Frankel, O.H. (1958) The dynamics of plant breeding. Journal of the Australian Institute of Agricultural Science. 24:112-123.

Fufa, H., Baenziger, P. S., Beecher, B. S., Dweikat, I., Graybosch, R. A. and Eskridge, K. M. (2005). Comparison of phenotypic and molecular marker-based classifications of hard red winter wheat cultivars. Euphytica. 145:133–146.

Gepts, P. and Papa, R. (2002) Evolution during domestication. In: Encyclopedia of Life Sciences. London: Nature Publishing Group.

Gepts, P. (2004) Crop domestication as a long-term selection experiment. Plant Breed Rev 24 (Part 2): 1-44

Gezan, S. A., White, T. L. and Huber, D. A. (2010). Accounting for Spatial Variability in Breeding Trials: A Simulation Study. Agronomy Journal. 102(6):1562.

Gianola, D. and van Kaam, J. B. C. H. M. (2008). Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 178(4):2289–303.

Goddard, M. E. and Hayes, B. J.(2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Review. Genetics. 10:381–391. Abstact.

Grabowski, P., Morris, G., Casler, M. and Borevitz, J.O. (2012) Range-wide Genomic Variation and Population Structure of Switchgrass (Panicum virgatum L.) Measured Using Genotyping-By-Sequencing (GBS). International Plant and Animal Genome Conference.[En línea]. https://pag.confex.com/pag/xx/webprogram/Paper1297.html

Grattapaglia, D. (2012). Twenty Years of Eucalyptus Molecular Breeding: From RAPD and QTLs to Genotyping-By-Sequencing and Genomic Selection of Complex Traits. International Plant and Animal Genome Conference.[En línea]. https://pag.confex.com/pag/xx/webprogram/Paper2695.html

Gu, Y.Q., Coleman-Derr, D., Kong, X.Y. and Anderson, O.D. (2004). Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four Triticeae genomes. Plant Physiol. 135:459-470.

Guo, X., Gao, A., Liu, W., Yang, X., Li, X. and Li, L. (2011). Evaluation of genetic diversity , population structure and linkage disequilibrium among elite Chinese wheat ( Triticum aestivum L .) cultivars. Australian Journal of Crop Science. 5(10):1167–1172.

Gupta, P. K., Rustgi, S. and Kulwal, P. L. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant molecular biology. 57(4):461–85.

Gupta, P. K., Rustgi, S. and Mir, R. R. (2008). Array-based high-throughput DNA markers for crop improvement. Heredity. 101(1):5–18.

Gustafson, D. I., Horak, M. J., Rempel, C. B., Metz, S. G., Gigax, D. R. and Hucl, P. (2005). An Empirical Model for Pollen-Mediated Gene Flow in Wheat. Crop Science. 45(4): 1286.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome biology. 10(3):R32.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. and Goddard, M.E. (2009) Genomic selection in dairy cattle: progress and challenges. Journal of Dairy Science. 92:433–443

Heisey, P.W., Lantican, M.A. and Dubin, H.J. (2002). Impacts of International Wheat Breeding Research in Developing

Heffner, E. L., Sorrells, M. E. and Jannink, J.L. (2009). Genomic Selection for Crop Improvement. Crop Science. 49(1):1-12.

Heffner, E. L., Lorenz, A. J., Jannink, J.-L. and Sorrells, M. E. (2010). Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Science. 50(5):1681.

Heffner, E. L., Jannink, J. and Sorrells, M. E. (2011). Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. The Plant Genome. 4:65–75.

Helyar, S. J., Limborg, M. T., Bekkevold, D., Babbucci, M., van Houdt, J., Maes, G. E., Bargelloni, L., et al. (2012). SNP discovery using Next Generation Transcriptomic Sequencing in Atlantic herring (Clupea harengus). PloS one. 7(8):e42089.

Hendre, P. S., Kamalakannan, R. and Varghese, M. (2012). High-throughput and parallel SNP discovery in selected candidate genes in Eucalyptus camaldulensis using Illumina NGS platform. Plant biotechnology journal. 10(6):646–56

Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., Fox, P., et al. (2008). Whole-genome sequencing and variant discovery in C . elegans. Nature Methods. 5(2):183-188.

Holland, J. B., Nyquist, W. E. and Cervantes-Martínez, C. T. (2003). Heritability for Plant Breeding: An Update. Plant Breeding Reviews. 22:9–112.

Hori, K., Kobayashi, T., Shimizu, A., Sato, K., Takeda, K. and Kawasak,i S. (2003) Efficient construction of high-density linkage map and its application to QTL analysis in barley. Theorical and Applied Genetics. 107:806–813

Huang, X., Paulo, M., Boer, M., Effgen, S., Keizer, P. and Koornneef, M. (2011). Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. PNAS. 108(11):4488–4493.

Hyten, D. L., Choi, I.-Y., Song, Q., Specht, J. E., Carter, T. E., Shoemaker, R. C., Hwang, E.-Y., et al. (2010a). A High Density Integrated Genetic Linkage Map of Soybean and the Development of a 1536 Universal Soy Linkage Panel for Quantitative Trait Locus Mapping. Crop Science. 50(3):960-968.

Hyten, D. L., Cannon, S. B., Song, Q., Weeks, N., Fickus, E. W., Shoemaker, R. C., Specht, J. E., et al. (2010b). High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC genomics. 11:38.

Hyten, D. L., Song, Q., Fickus, E. W., Quigley, C. V., Lim, J.-S., Choi, I.-Y., Hwang, E.-Y., et al. (2010c). High-throughput SNP discovery and assay development in common bean. BMC genomics. 11:475.

Ingvarsson, P. K. and Street, N. R. (2011). Association genetics of complex traits in plants. The New phytologist. 189(4):909–22.

Iwata, H. and Jannink, J.-L. (2011). Accuracy of Genomic Selection Prediction in Barley Breeding Programs: A Simulation Study Based On the Real Single Nucleotide Polymorphism Data of Barley Breeding Lines. Crop Science. 51(5):1915.

Jaccoud, D., Peng, K., Feinstein, D. and Kilian, A. (2001). Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic acids research, 29(4), E25.

Jannink, J. L. and Walsh, J. B. (2002) Association mapping in plant populations, pp. 59–68 in Quantitative Genetics Genomics and Plant Breeding, edited by M.S. Kang. CABI, Wallingford, UK

Jorde, L.B. 2000. Linkage disequilibrium and the search for complex disease genes. Genome Research. 10: 1435-1444.

Kang, H. M., Zaitlen, N. a, Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J. and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics. 178(3):1709–23.

Kearsey, M.J. and Farquhar, A.G.L. (1998) QTL analysis; where are we now? Heredity. 80(2):137-142

Khush, G. S. (2001). Green revolution: the way forward. Nature reviews. Genetics. 2(10):815–22.

Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., et al. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. BMC bioinformatics. 12(1):231.

Kirigwi, F. M., Ginkel, M., Brown-Guedira, G., Gill, B. S., Paulsen, G. M. and Fritz, a. K. (2007). Markers associated with a QTL for grain yield in wheat under drought. Molecular Breeding. 20(4):401–413.

Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M. and Seelow, D. (2011). Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PloS one. 6(11):e28240.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K. and Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics (Oxford, England). 25 (17):2283-2285

Koebner, R. M. D. and Summers, R. W. (2003). 21St Century Wheat Breeding: Plot Selection or Plate Detection? Trends in biotechnology. 21(2):59–63.

Lafever, H.N. and Patterson, F.L. (1965) Tests for Association of Genetic Factors in a Nullisomic Line of Avena sativa. Crp Science. 5(6):509-510.

Lande, R. and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics. 124(3):743–56. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1203965andtool=pmcentrezandrendertype=abstract [Verified: 5 Nov 2012].

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., et al. (2001). Initial sequencing and analysis of the human genome. Nature. 409(6822):860-921.

Legarra, A., Robert-Granié, C., Manfredi, E. and Elsen, J.M. (2008). Performance of genomic selection in mice. Genetics. 180(1):611–8.

Leiser, W. L., Rattunde, H. F., Piepho, H.-P. and Parzies, H. K. (2012). Getting the Most Out of Sorghum Low-Input Field Trials in West Africa Using Spatial Adjustment. Journal of Agronomy and Crop Science. 198(5):349–359.

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G. and Webb, W. W. (2003) Zero-Mode Waveguides for Singles-Molecule Analysis at High Concentrations. Science. 299 (5607):682-686.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 25:2078-2079.

Li, H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754-60.

Lu, X., Wei, S., Zhang, Y. and Lu, H. (2011). Epistatic Association Mapping in Homozygous Crop Cultivars. PloS one. 6(3):e17773.

Mackay, T. F. C., Stone, E. a and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. Nature reviews. Genetics. 10(8):565–77.

Manifesto, M. M., Schlatter, A. R., Hopp, H. E., Sua, E. Y. and Dubcovsky, J. (2000). Quantitative Evaluation of Genetic Diversity in Wheat Germplasm. Crop Science. 4:682–690.

Martin, T. J. (1990) Outcrossing in twelve hard red winter wheat cultivars. Crop Science 30(1):59–62

McFadden, E. S. and Sears, E. R. (1946) The origin of Triticum spelta and its free-threshing hexaploid relatives. Journal of Heredity. 37:81–89.

McKenna, A., Hanna, M., Banks, E., Sivachenko., A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 20:1297-303.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome research. 19(9):1527–41.

McNeish, R. S. (1992). The Origins of Agriculture and Settle Life. Univesity of Oklahoma. Press, Norman and London.

Metzker, M. L. (2010). Sequencing technologies - the next generation. Nature reviews. Genetics. 11(1):31-46.

Meuwissen, T. H., Hayes, B. J. and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157(4):1819-29.

Müller, B. U., Schützenmeister, A. and Piepho, H.-P. (2010). Arrangement of check plots in augmented block designs when spatial analysis is used. Plant Breeding. 129(6):581–589.

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E. and Buckler, E. S. (2009). Association mapping: critical considerations shift from genotyping to experimental design. The Plant cell. 21(8):2194–202.

Nakaya, A. and Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? Annals of botany. Disponible en www.aob.oxfordjournals.org (verificado en 22 Oct. 2012)

Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P. and Barron, A. E. (2011). Landscape of next-generation sequencing technologies. Analytical chemistry. 83(12):4327–4341.

Nielsen, R., Paul, J. S., Albrechtsen, A. and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. Nature reviews. Genetics. 12(6):443-51.

Ordas, B., Malvar, R. a, Santiago, R. and Butron, A. (2010). QTL mapping for Mediterranean corn borer resistance in European flint germplasm using recombinant inbred lines. BMC genomics. 11:174.

Prakash, C. S. (2001). The genetically modified crop debate in the context of agricultural evolution. Plant physiology. 126(1):8–15.

Papadakis, J. S. (1937). Methode statistique pour des experiences sur champ. Bulletin de l'Institut d'Amelioration des Plantes a' Salonique. 23.

Paradis, E., Claude, J. and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 20:289-290.

Parisseaux B. and Bernardo R. (2004) In silico mapping of quantitative trait loci in maize. Theorical and Applied Genetics. 109:508-514

Paterson, A. H., Lander E. S., Hewitt J. D., Peterson S., Lincoln S. E. and Tanksley S. D., (1988). Resolution of quantitative traits into Mendelian factors, using a complete linkage map of restriction fragment length polymorphisms. Nature. 335:721-726.

Peiris, T. U. S., Samita, S. and Veronica, W. H. D. (2008). Accounting for Spatial Variability in Field Experiments on Tea. Experimental Agriculture. 44(04):547.

Peterson, R.K.D. and Shama, L.M. (2005). A Comparative Risk Assessment of Genetically Engineered, Mutagenic and Conventional Wheat Production Systems. Biomediacl and Life Sciences. 14(6):859-875.

Piepho, H. P. and Williams, E. R. (2010). Linear variance models for plant breeding trials. Plant Breeding. 129(1):1–8.

Poland, J. A., Brown, P. J., Sorrells, M. E. and Jannink, J.-L. (2012a). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PloS one. 7(2):e32253.

Poland, J.A. and Rife, T.W. (2012b). Genotyping by sequencing for plant breeding. The Plant Genome. Aceptado

Poland, J.A., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sanchez-Villeda, H., Sorrells, M. and Jannink, J.L. (2012c). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. (In Press)

Ratan, A., Zhang, Y., Hayes, V. M., Schuster, S. C. and Miller, W. (2010). Calling SNPs without a reference sequence. BMC bioinformatics. 11:130.

Resende, M. F. R., Muñoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., Resende, M. D. V., et al. (2012). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. The New phytologist. 193(3):617–24.

Rieseberg, L. H. and Carney, S. E. (1998). Plant hybridization. New Phytology, 140:599–624.

Robertson, A. (1967). The nature of quantitative genetic variation. In Heritage from Mendel (ed. R. A. Brink), pp. 265-280. Madison: University of Wisconsin Press.

Sarkar P and Stebbins GL (1956) Morphological evidence concerning the B genome in wheat. Am J Bot 43(4), 297–304.

Sanger, F. and Nicklen, S. (1977). DNA sequencing with chain-terminating. Proceedings of the National Academy of Sciences of the United States of America, 74(12), 5463-5467.

Schnurbusch, T.H., Huang, C.,Collins N.C., Sutton T., John U., Roy S., Paltridge N., Tester M., Paltridge P. and Fincher G.B. (2007) Prospects for developing GM varieties of wheat and barley with enhanced tolerance to abiotic stress. Agricultural Science. 21:4-10

Schwender, H., Li, Q., Neumann C. and Ruczinski, I. (2012) trio: trio package whithout Fortran code. R package version 1.7.0. http://CRAN.R-project.org/package=trio

Shin, J. H., Blay, S., McNeney, B. and Graham. (2006) LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. J Stat Soft. 16:Code Snippet 36

Smith, L.I. (2002) A tutorial on Principal Components Analysis. http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Somers D.J., Kirkpatrick R., Moniwa M., Walsh A. (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome. 49: 431–437.

Sommers D.J., Banks T., Depauw R, Fox S., Clarke J., Pozniak C. andMcCartney C. (2007) Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat. Genome. 50(6):557-567.

Sorrells M.E. (1998). Marker Assisted Selection: Is it Practical ?. In: International Workshop on the Aplication of Biotechnologies to Wheat Breeding. Eds: Dr. Mohan Kolhi and INIA La Estanzuela

Stuber, C.W., Lincoln, S.E., Wolff, D.W., Helentjaris, T. and Lander, E. S. (1992). Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics. 132(3):823–39.

Stuber, C.W., Polacco, M. and Senior M.L. (1999) Synergy of Empirical Breeding, Marker-Assisted Selection and Genomics to Increase Crop Yield Potential. Crop Sience. 39(6):1571-1583.

Tanksley, S.D. and J.C. Nelson (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theorical and Applied Genetics. 92:191-203.

Technow, F., (2011) Rpackage mvngGrAd: moving grid adjustment in plant breeding field trials. R packaage version 0.1.2.

Trebbi, D, Maccaferri, M., de Heer P., Sørensen A., Giuliani S., Salvi S., Sanguineti M. C., Massi A.,et al. (2011) High-throughput SNP discovery and genotyping in durum wheat (Triticum durum Desf.). Theorical and Applied Genetics. 123(4):555-569.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. Journal of dairy science. *91*(11):4414–23.

Varala, K., Swaminathan, K., Li, Y. and Hudson, M. E. (2011). Rapid genotyping of soybean cultivars using high throughput sequencing. PloS one, 6(9), e24811.

Varshney, R.K., Graner, A., Sorrells, M.E. (2005) Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 23(1):48–55

Vaughan, D. a, Balázs, E. and Heslop-Harrison, J. S. (2007). From crop domestication to super-domestication. Annals of botany. 100(5):893–901.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., et al. (2001). The sequence of the human genome. Science (New York, N.Y.), 291(5507), 1304-51.

Verhoeven, K. J. F., Casella, G. and McIntyre, L. M. (2010). Epistasis: obstacle or advantage for mapping complex traits? PloS one. 5(8):e12264.

Von Zitzewitz, J., Cuesta-Marcos, A., Condon, F., Castro, A. J., Chao, S., Corey, A., Filichkin, T., et al. (2011). The Genetics of Winterhardiness in Barley: Perspectives from Genome-Wide Association Mapping. The Plant Genome Journal. 4(1):76.

Wall, J. D. and Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. Nature reviews. Genetics. 4(8):587–97.

Wray, N. and Visscher, P. (2008) Estimating trait heritability. Nature Education. 1:1-16

Wu, X., Ren, C., Joshi, T., Vuong, T., Xu, D. and Nguyen, H. T. (2010). SNP discovery by high-throughput sequencing in soybean. BMC genomics, 11:469.

Wu, X., Chang, X. and Jing, R. (2012). Genetic insight into yield-associated traits of wheat grown in multiple rain-fed environments. PloS one. 7(2):e31249.

Xie, C. X., Warburton, M., Li, M. S., Li, X. H., Xiao, M. J., Hao, Z. F., Zhao, Q. and Zhang, S. H. (2008). An analysis of population structure and linkage disequilibrium using multilocus data in 187 maize inbred lines. Molecular Breeding. 21: 407–418

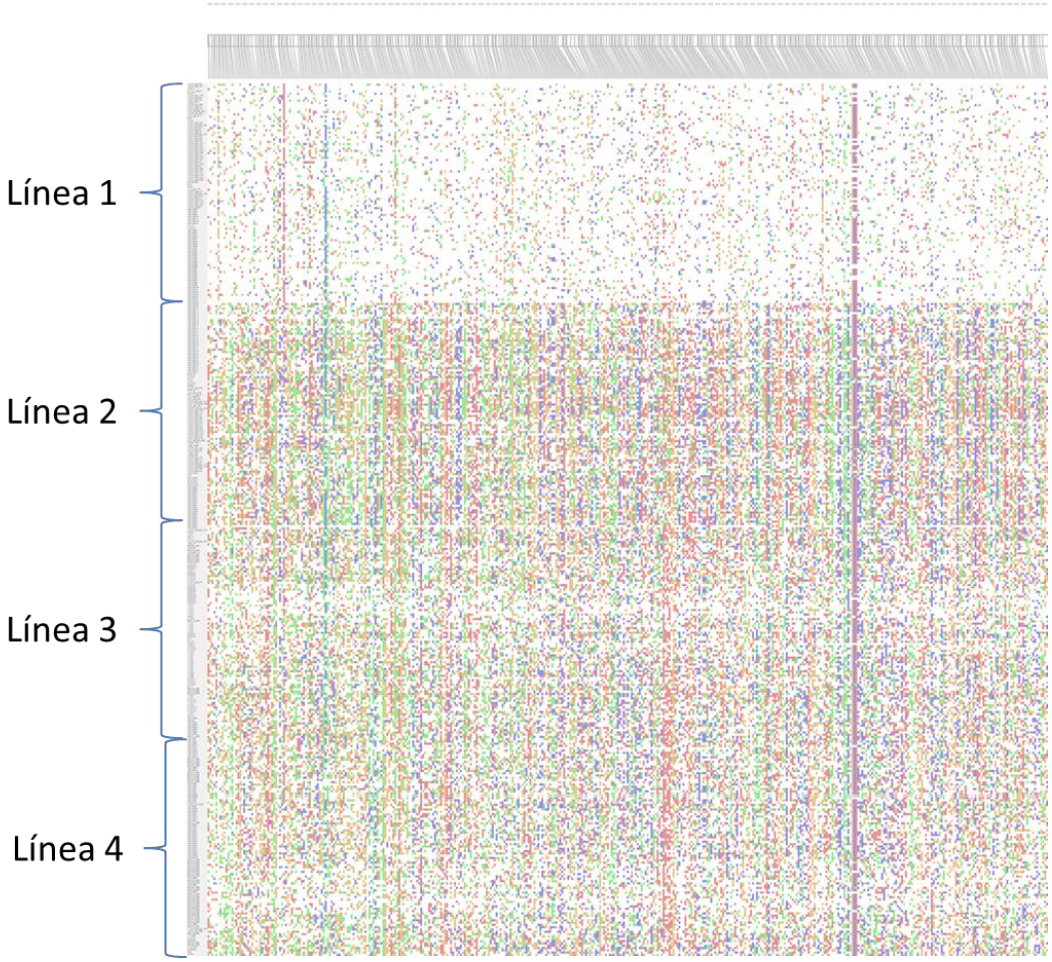
Yang, J., Benyamin, B., McEvoy B. P., Gordon, S., Henders, A.K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. and Visscher, P. M. (2010) Common SNPs explain a large proportion of the heritability for human height. Nature Gen. 42: 565–569.

You, F. M., Huo, N., Deal, K. R., Gu, Y. Q., Luo, M.-C., McGuire, P. E., Dvorák, J., et al. (2011). Annotation-based genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. BMC genomics. 12(1):59.

Zamir, D. (2001). Improving plant breeding with exotic genetic libraries. Nature reviews. Genetics. 2(12):983–9.

Zhang, L., Liu, D., Guo, X., Yang, W., Sun, J., Wang, D., Sourdille, P., et al. (2011). Investigation of genetic diversity and population structure of common wheat cultivars in northern China using DArT markers. BMC genetics. 12(1):42.

## 7. ANEXOS




Anexo 1. Para cada muestra (filas) se indica el alelo presente en el SNP (columnas) con un código de colores para cada base nucleotídica. Las muestras se encuentran ordenadas según las líneas del secuenciador.

| Línea 1 | | Línea 2 | | Línea 3 | | Línea 4 | |
|---|---|---|---|---|---|---|---|
| 2 | :1099985 | 2 | :1007022 | 2 | :1038585 | 2 | :1079689 |
| 3 | : 351294 | 3 | : 453369 | 3 | : 442689 | 3 | : 456116 |
| 4 | : 184469 | 4 | : 294158 | 4 | : 288640 | 4 | : 292512 |
| 5 | : 124432 | 5 | : 218001 | 5 | : 217964 | 5 | : 217847 |
| 6 | : 95084 | 6 | : 172267 | 6 | : 172446 | 6 | : 171944 |
| 7 | : 78079 | 7 | : 139382 | 7 | : 140458 | 7 | : 140488 |
| 8 | : 66654 | 8 | : 115244 | 8 | : 116618 | 8 | : 116415 |
| 9 | : 60191 | 9 | : 97530 | 9 | : 99394 | 9 | : 99032 |
| 10 | : 54078 | 10 | : 84341 | 10 | : 84793 | 10 | : 85641 |
| 11 | : 49764 | 11 | : 72861 | 11 | : 74595 | 11 | : 74674 |
| 12 | : 46565 | 12 | : 64342 | 12 | : 66013 | 12 | : 66644 |
| 13 | : 44587 | 13 | : 58084 | 13 | : 58820 | 13 | : 59949 |
| 14 | : 41908 | 14 | : 52316 | 14 | : 53409 | 14 | : 54085 |
| 15 | : 39938 | 15 | : 47713 | 15 | : 48825 | 15 | : 48713 |
| 16 | : 38361 | 16 | : 43515 | 16 | : 44674 | 16 | : 44676 |
| 17 | : 37154 | 17 | : 39788 | 17 | : 40747 | 17 | : 41119 |
| 18 | : 35176 | 18 | : 36656 | 18 | : 37552 | 18 | : 37733 |
| 19 | : 33483 | 19 | : 33668 | 19 | : 34751 | 19 | : 35057 |
| 20 | : 32606 | 20 | : 31225 | 20 | : 32301 | 20 | : 32696 |
| 21 | : 31142 | 21 | : 29197 | 21 | : 30089 | 21 | : 30510 |
| 22 | : 29167 | 22 | : 27373 | 22 | : 28116 | 22 | : 28028 |
| 23 | : 27723 | 23 | : 24966 | 23 | : 26165 | 23 | : 26236 |
| 24 | : 25783 | 24 | : 23642 | 24 | : 24344 | 24 | : 24375 |
| 25 | : 24057 | 25 | : 22479 | 25 | : 23213 | 25 | : 23168 |

| Línea 1 | | Línea 2 | | Línea 3 | | Línea 4 | |
|---|---|---|---|---|---|---|---|
| 26 | : 22397 | 26 | : 21004 | 26 | : 21694 | 26 | : 21843 |
| 27 | : 20788 | 27 | : 19632 | 27 | : 20298 | 27 | : 20438 |
| 28 | : 19163 | 28 | : 18718 | 28 | : 19124 | 28 | : 19397 |
| 29 | : 17732 | 29 | : 17671 | 29 | : 18203 | 29 | : 18232 |
| 30 | : 16747 | 30 | : 16775 | 30 | : 17211 | 30 | : 17240 |
| 31 | : 15288 | 31 | : 15948 | 31 | : 16417 | 31 | : 16399 |
| 32 | : 14314 | 32 | : 15137 | 32 | : 15232 | 32 | : 15606 |
| 33 | : 13469 | 33 | : 14460 | 33 | : 14752 | 33 | : 14675 |
| 34 | : 12711 | 34 | : 13932 | 34 | : 14072 | 34 | : 14007 |
| 35 | : 11901 | 35 | : 13009 | 35 | : 13406 | 35 | : 13181 |
| 36 | : 11669 | 36 | : 12566 | 36 | : 13015 | 36 | : 12776 |
| 37 | : 11259 | 37 | : 11949 | 37 | : 12339 | 37 | : 12245 |
| 38 | : 10651 | 38 | : 11366 | 38 | : 11688 | 38 | : 11679 |
| 39 | : 10259 | 39 | : 10966 | 39 | : 11234 | 39 | : 11017 |
| 40 | : 9840 | 40 | : 10761 | 40 | : 10856 | 40 | : 10537 |
| 41 | : 9447 | 41 | : 9975 | 41 | : 10559 | 41 | : 10117 |
| 42 | : 9262 | 42 | : 9628 | 42 | : 10018 | 42 | : 9807 |
| 43 | : 8910 | 43 | : 9310 | 43 | : 9538 | 43 | : 9257 |
| 44 | : 8780 | 44 | : 8876 | 44 | : 9234 | 44 | : 9053 |
| 45 | : 8384 | 45 | : 8648 | 45 | : 9006 | 45 | : 8657 |
| 46 | : 8095 | 46 | : 8142 | 46 | : 8555 | 46 | : 8417 |
| 47 | : 7824 | 47 | : 7993 | 47 | : 8272 | 47 | : 8108 |
| 48 | : 7706 | 48 | : 7768 | 48 | : 7959 | 48 | : 7782 |
| 49 | : 7388 | 49 | : 7427 | 49 | : 7736 | 49 | : 7514 |
| 50 | : 7209 | 50 | : 7282 | 50 | : 7399 | 50 | : 7172 |
| (Other): 407559 | | (Other): 439272 | | (Other): 450504 | | (Other): 435423 | |

Anexo 2. Se muestra la profundidad de la secuencias en cada una de la líneas del secuenciador. Se muestra hasta profundidades de 50, luego se juntan lo valores para profundidades más altas.