UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE AGRONOMÍA

OPTIMIZACIÓN DE LA POBLACIÓN DE ENTRENAMIENTO PARA SELECCIÓN GENÓMICA EN TRIGO Y ARROZ

por

María Inés BERRO ROVELLA

TESIS presentada como uno de los requisitos para obtener el título de *Magister* en Ciencias Agrarias opción Bioestadística

Montevideo
URUGUAY
Mayo 2017

Tesis aprobada por el tribunal integrado por Lic. en Biología (Ph.D.) Francisco Peñagaricano, Ing. Agr. Ph.D. Martín Quincke y Ph.D. Pablo Speranza, el 15 de Mayo de 2017. Autor/a: Lic. en Estadística María Inés Berro. Directora Ing. Agr. (Ph.D.) Lucía Gutiérrez.

AGRADECIMIENTOS

Quiero agradecer a todos los que participaron de una u otra manera durante el proceso de este trabajo. En particular y primer lugar, a mi tutora Lucía Gutiérrez, por su guía, apoyo, dedicación y generosidad en los conocimientos y experiencias trasmitidas. A mis compañeros del Departamento de Biometría, Estadística y Computación (DBEC) de Facultad de Agronomía, en especial a Bettina, Ale, Andrea, Pablo, Natalia, Virginia, Rafael, Víctor, Pablo y Ana Lía gracias por los aportes, charlas y consejos. A Martín Quincke, Francisco Peñagaricano y Pablo Esperanza por sus valiosas y fundamentales aportes en su rol de tribunal de este trabajo. A mi familia, a mis padres por apoyar incondicionalmente y en especial a Azul y Julia que hacen que todo valga más la pena. A Kiwi por estar siempre e ir juntos a la par.

TABLA DE CONTENIDO

Página

RESUMEN

La Selección Genómica (GS) busca mejorar la precisión de la selección de individuos a través de la estimación de los valores de cría genéticos (GEBV) usando información de marcadores moleculares. A través de un modelo estadístico estimado con información fenotípica y genotípica de una población de entrenamiento (PE) se predicen los GEBVs de individuos que solo cuentan con información genotípica. La exactitud de las predicciones es afectada por el modelo de predicción, por el número y tipo de marcadores moleculares, por la arquitectura del carácter y por el número y la estructura de los individuos de la PE y su relación con la población a predecir (PP). Este trabajo tiene como objetivos (1) evaluar cómo afecta el tamaño y el relacionamiento genético entre los individuos de la PE en la exactitud de las predicciones y (2) analizar cómo el relacionamiento genético entre la PP y PE afecta la exactitud de las predicciones. Se usó una población de trigo compuesta de 1353 líneas avanzadas genotipadas con 81999 SNPs evaluada para rendimiento en grano y una población de arroz con 644 líneas avanzadas genotipadas con 15000 SNPs y evaluada para rendimiento en grano. Se ajustaron modelos G-BLUP según tres estrategias de agrupamientos: basada en el relacionamiento genético para las dos poblaciones, según los ensayos del programa de mejoramiento y según el ciclo de las líneas para la población de trigo. Se emplearon dos estrategias para elegir líneas de la PE para predecir PP: (a) según la matriz de relacionamiento genético con la PP y (b) según la matriz de similitud genética ponderada por el efecto estimado de los marcadores moleculares. La exactitud se midió con el coeficiente de correlación lineal de Pearson entre los valores observados del carácter y los valores predichos. Se encontraron tres grupos que presentaron mayor exactitud en las predicciones que con grupos del mismo tamaño pero elegidos al azar y con la utilización toda la PE. Modelar la estructura según la información de ensayos y ciclo del cultivo mejora las precisiones. Incluir la estimación de los efectos de los marcadores moleculares mejoró la precisión respecto de utilizar solo el relacionamiento genético.

Palabras clave: selección genómica, exactitud, población de entrenamiento.

TRAINING POPULATION OPTIMIZATION FOR GENOMIC SELECTION

ABSTRACT

Improving the accuracy of line selection through the estimation of the genetic values (GEBV) for characters using information of molecular markers is one of the goals of genomic selection (GS). A prediction model estimated from genomic and phenotypic information of a training population allows the calculation of the estimation of GEBV for lines that only have genotypic information. The accuracy of the predictions of the GS in the programs is affected by the prediction model, the number and type of molecular markers, the architecture of the character, and the size and structure of the training population (TR) and its relationship with the population to predict (TE). The objectives of this study were: (1) to evaluate how the size and the genetic relationships between the individuals in the training population affect the prediction accuracy; (2) to evaluate the effect of the genetic relationship between the TR and TE in prediction accuracy. A total of 1353 advanced lines of wheat genotyped with 81999 genotyping-by sequencing (GBS) markers, and a population of 644 advanced lines of rice with 15000 SNPs were evaluated for yield grain. G-BLUP models were adjusted according to three strategies of grouping: based in the genetic relationship for two populations, according to the trials of the program and according to the cycle of the lines for the wheat population. Two strategies were used to choose lines of the TR to predict TE: (a) according to a matrix of genetic similarity with the TE, and b) according to the effect of the molecular markers. Accuracy was measured with Pearson's linear correlation coefficient between the observed and predicted values. Three groups were found that presented greater accuracy than other groups of the same size, selected at random and also among the population in general. Modeling the structure using trial information and the crop cycle improved the accuracy. The addition of the estimation of the effects of the molecular markers improved the accuracy compared to the single use of the genetic relationship.

Keywords: genomic selection, accuracy, training population

1. INTRODUCCIÓN

El trigo y el arroz son cultivos de importancia mundial en términos económicos y productivos por ser pilares básicos en la alimentación humana (Hawkesford et al. 2013). Actualmente la producción mundial de trigo es de 718 millones de toneladas y la producción de arroz ronda las 496,6 millones de toneladas (FAO, 2014). En Uruguay, el trigo es el principal cultivo de invierno, con una producción de 982.4 miles de toneladas en 2012/13 y un rendimiento promedio de 2.183 toneladas por hectárea (DIEA-MGAP, 2014). Por otro lado, Uruguay es el principal exportador de arroz de América Latina, con una producción de 1359,5 miles de toneladas en 2012/13 y un rendimiento promedio de 7,88 toneladas por hectárea (DIEA-MGAP, 2014).

El mejoramiento genético ha permitido mejorar la productividad de estos cultivos de forma sostenida, no obstante, se han alcanzado límites en la ganancia genética con los métodos tradicionales (Lande y Thompson, 1990). Las nuevas tecnologías con marcadores moleculares hacen cada vez más accesible el genotipado de todo el genoma de cultivos con la rapidez necesaria para utilizarlo en los programas de mejoramiento (Bonnet et al., 2005; Bernardo, 2008). La introducción de selección basada en información molecular permite acortar el tiempo entre el cruzamiento y la obtención de las variedades (Lorenz et al. 2011, Bernardo y Yu, 2007).

La Selección Genómica (GS) tiene como objetivo mejorar la precisión de la selección de líneas a través de la estimación de los valores de cría genéticos (GEBV) para caracteres fenotípicos usando información de marcadores moleculares (Meuwissen et al., 2001, Isidro et al., 2015). La GS explota el desequilibrio de ligamiento (LD) (el cual hace referencia a la relación existente entre las frecuencias alélicas o haplotípicas observadas en una posición del genoma frente a otra) (de Roos et al, 2008). La GS como metodología de selección tiene ventajas respecto a otras aproximaciones genéticas ya que incluye una gran cantidad de marcadores moleculares en la predicción considerando QTL (loci de un caracteres cuantitativos) de efectos mayores, pero también de efectos menores (Xu, 2003; Poland y Rife,

2012; Bernardo y Yu, 2007; Lorenz, 2011). A través de un modelo estadístico estimado a partir de información fenotípica y genotípica de una muestra de entrenamiento (población de entrenamiento) se predicen los GEBVs de líneas que solo cuentan con información genotípica (Meuwissen et al., 2001). Independientemente del tipo de población con la que se trabaje, la implementación de GS se puede resumir en cuatro pasos: (i) diseñar poblaciones de entrenamiento con datos fenotípicos y genotípicos, (ii) estimar los efectos de los marcadores en la población de entrenamiento, (iii) predecir los GEBV de nuevas líneas con datos genotípicos, y (iv) realizar la selección (Heffner et al., 2009; Jannink et al., 2010).

Datos simulados en plantas muestran que la GS podría acelerar el progreso en mejoramiento genético vegetal (Heslot et al, 2012), obteniéndose mayores ganancias que con selección asistida; mayor ganancia por unidad de costo, y mayor capacidad para generar mejores progenies (Poland y Rife, 2012). Por otro lado, en los casos donde la información fenotípica no está disponible o no es confiable se espera que la aplicación de GS en combinación con otras estrategias en los programas permita acelerar los ciclos de mejoramiento (Lorenzana y Bernardo, 2009; Heffner 2009, 2011; Hickey et al, 2015).

La precisión de las predicciones de la GS, se ve afectada por varios elementos: (1) el modelo de predicción empleado (Lorenzana y Bernardo, 2009), (2) el número y tipo de marcadores moleculares (Lorenzana y Bernardo, 2009; Asoro et al., 2011), (3) la arquitectura del carácter (Combs y Bernardo, 2013) y (4) el número y la estructura de las líneas de la población de entrenamiento y su relación con la población a predecir (de los Campos, 2012; Isidro et al., 2015). Muchos estudios de SG se han focalizado en el desarrollo y la evaluación de distintos modelos estadísticos que se diferencian principalmente por los supuestos distribucionales de los efectos de los marcadores. Entre ellos se encuentran modelos de regresión que establecen relaciones aditivas entre marcadores, realizando una contracción en la estimación de los efectos para resolver el problema de la sobreparametrización y asumiendo varianzas homogéneas, como ser Ridge Regression (Meuwissen et al., 2001) y el G-BLUP (de los Campos el al., 2012). Además, se desarrollaron modelos Bayesianos en los que se asumen

diferentes distribuciones a priori para las varianzas de los marcadores (Meuwissen et al. 2001, Habier et al., 2007), métodos no paramétricos y redes neuronales (Gianola et al., 2011). Pero varios trabajos muestran que las ganancias en la precisión de las predicciones al considerar diferentes modelos son marginales (Asoro et al., 2011, Heffner et al., 2011). La densidad de cobertura de los marcadores o el número de los marcadores, que son utilizados como covariables en los modelos de predicción no sería una limitante ya que con los avances tecnológicos es posible tener muy buena densidad y cobertura de marcadores a lo largo del genoma (Asoro et al., 2011; Lorenzana y Bernardo, 2009). Sin embargo, es sabido que el número de marcadores óptimo, el que arroja mayor precisión en las predicciones, depende del tipo de población y del carácter (Combs y Bernardo, 2013). En general, a medida que se incrementa el número de marcadores crece la precisión de las predicciones hasta un cierto límite o techo donde se deja de crecer. La magnitud de este límite depende del tipo de carácter y de la naturaleza de la población (Lorenzana y Bernardo, 2009; Heffner et al., 2011; Asoro et al., 2011). A mayor heredabilidad más alta es la precisión de las predicciones, no obstante, a igual heredabilidad caracteres que se ven afectados por pocos genes (oligogénicos) presentan mejores precisiones en las predicciones de los valores de cría genético que los caracteres afectados por múltiples genes (poligénicos) (Combs y Bernardo, 2013; Isidro et al., 2015).

Finalmente, uno de los factores más importantes a tener en cuenta es la composición de la población de entrenamiento (Lorenz et al., 2012, Reidelsheimer et al., 2013). Encontrar el conjunto de líneas más informativo para entrenar el modelo de predicción es un objetivo clave en SG. A mayor número de líneas incluidas en el entrenamiento del modelo mejores son las precisiones (Meuwissen, 2009), no obstante, esta relación no es lineal ni el incremento es constante para todos los tamaños de la población de entrenamiento (Lorenzana y Bernardo, 2009; Asoro et al., 2011; Clark et al., 2012; Lorenz et al.; 2012, Wientjes et al.; 2013, Pszczola et al., 2012; Habier et al., 2007). Se ha observado que a partir de un cierto número de líneas los incrementos en la precisión comienzan a ser marginales hasta que llegan a un límite donde el incremento deja de ser significativo (Asoro et al., 2011). Por otro lado, el mayor relacionamiento entre las líneas de la población de entrenamiento

lleva a que las precisiones de las predicciones sean mayores (Habier et al., 2007). Cuanta más estructura en la población, al entrenar modelos por bloques o grupos de líneas más homogéneas se obtienen mayores precisiones en las predicciones (Asoro et al., 2011), aun con tamaños de población menores. El relacionamiento con la población a predecir determina la precisión de los modelos, si la población a predecir tiene un alto relacionamiento con la de entrenamiento las predicciones de los modelos serán más precisas (Crossa et al., 2014; Clark et al., 2012; Lorenz et al., 2012; Wientjes et al., 2013; Pszczola et al., 2012; Habier et al., 2007).

No está claro aún cuál es la mejor estrategia para la construcción de la población de entrenamiento, resultados muestran que las estrategias óptimas varían según el factor o el grupo de factores que están presentes en las poblaciones de entrenamiento (Pszczola et al., 2012; Reidelsheimer et al., 2013). El objetivo de este trabajo es evaluar y analizar estrategias para optimizar la población de entrenamiento para los programas de Mejoramiento Genético de Trigo y Arroz del Uruguay.

1.1 HIPÓTESIS Y OBJETIVOS

Hipótesis:

1. Dado un tamaño de muestra, la precisión de la predicción de modelos entrenados con líneas más relacionadas genéticamente es mayor que con líneas tomadas al azar.
2. Es posible encontrar algún subgrupo de la población de entrenamiento con el que se obtengan mayores precisiones en las predicciones que utilizando toda la población.
3. La mayor precisión en las predicciones se encontrará utilizando todas las líneas de la población de entrenamiento y modelando su relacionamiento o estructura genética.
4. Cuanto mayor sea la relación entre la población de entrenamiento y la población a predecir, mayor será la precisión de las predicciones; se puede encontrar una población de entrenamiento óptima según la población a predecir.

El objetivo general de este trabajo es evaluar y analizar estrategias de optimización de la población de entrenamiento para modelos de predicción genómica en los programas de mejoramiento genético de Arroz y Trigo del Uruguay.

Objetivos específicos:

1. Evaluar cómo afecta el tamaño, el relacionamiento genético entre las líneas y la estructura genética en sub-especies de la población de entrenamiento en la precisión de las predicciones.
2. Analizar cómo se afecta la precisión de las predicciones según el relacionamiento entre la población de entrenamiento y una población a predecir.

2. OPTIMIZATION OF TRAINING POPULATION FOR GENOMIC SELECTION[1]

2.1. SUMMARY

Genomic Selection (GS) use molecular markers to predict the breeding value of individuals using statistical models. The effectiveness of the GS in breeding programs depends on the genetic architecture traits, the prediction model, the number and type of markers and the structure of the training population. Our objectives were (1) evaluate the effect of the relationship of the lines, the genetic structure, and the size of the training population in the prediction accuracy using a single population for GS and (2) evaluate the effect of the genetic relationship between the training and testing populations in prediction accuracy. A total of 1535 advanced lines of wheat genotyped with 81999 SNPs and two subspecies of rice with 644 advanced lines, with three genotyped data set: Indica (40000 SNPs), Tropical Japonica (28000 SNPs) and combined Indica and Topical Japonica (15000 SNPs), both population were evaluated for grain yield. Using G-BLUP models Training and Testing Population using three strategies of grouping based: on genetic relationship, on crossing block historical relationship and on population structure, with subspecies. Two strategies were used to choose lines of the training population (TR) to predict a new population (TE): (1) choose lines based on their genetic similarity calculated with relationship matrix using three different criteria and (2) choose lines based on their markers scores weighted by their markers effect. We found some subsets of the population with which we obtained better accuracies: best for a random sample of the same size and better to use the entire population. Using the effect of the markers for calculating the distance was better than using genetic relationship matrix. The selection of individuals by the criterion of maximum value of relationship turned out to be good.

[1] Artículo a publicar en: The plant Genome

2.3. INTRODUCTION

One of the objectives of Genomic selection (GS) is improvement of the accuracy of the selection of lines through the estimation of the genetic values (GEBV) for characters using information of molecular markers **(**Meuwissen et al., 2001; Isidro, et al., 2015). Genomic and phenotypic information of a training population in a prediction model are used to estimate GEBV for lines that only have genotypic information (Meuwissen et al., 2001).

The GS as a selection methodology has advantages over other genetic approaches since it includes all molecular markers in the prediction considering QTL (quantitative trait loci) of major effects but also of smaller effects (Xu, 2003). Simulated plant data shows that it could accelerate progress in plant genetic improvement (Heslot et al, 2012), resulting in higher gains than using marker assisted selection; higher profit per unit of cost, and better progenies (Poland and Rife, 2012). On the other hand, in cases where phenotypic information is not available or is not reliable, it is expected that the application of GS in combination with other strategies in the programs will accelerate breeding cycles.

Many factors affect the accuracy of prediction including model, molecular marker density, training population composition, phenotyping traits, and phenotyping quality (Combs and Bernardo, 2013; Lorenz et al., 2012; Lorenz, 2013; Heffner et al., 2011). One of the most important factors is the composition of the training population (Lorenz et al., 2012; Reidelsheimer et al., 2013).

Several studies have demonstrated the importance of the genetic relationships between the training population individuals (TR) and the populations to predict (TE) in prediction accuracy (Clark et al., 2012; Lorenz et al., 2012; Wientjes et al., 2013; Pszczola et al., 2012; Habier et al., 2013). When the lines of the TR and TE are composed of germplasm from different breeding programs (Lorenz et al., 2012) or even among different families of complete siblings (Riedelsheimer et al., 2013) the prediction accuracy has turned out to be low. An explanation for these results is that the response to genomic selection is based on linkage desequilibrium (LD) binding specific alleles of SNPs and QTL (Meuwissen et al., 2001). The higher the LD, the

higher the accuracy predictions (Solberg et al., 2008). As the linkage disequilibrium between QTL and SNPs decreases after generations, the accuracy of genomic prediction is expected to decrease (Muir, 2007).

The genetic relationship between the training population and the population to be predicted is one of the most important elements to be included at the time of the choice of the TR composition.

The objective of this study was to compare strategies for optimizing the training population for genomic prediction models in the Rice and Wheat Breeding programs of Uruguay. First, we evaluated how size, genetic relationships structure in sub-species of the training population affect the accuracy of the predictions within the training population. Second, we studied how the accuracy of predictions according to the relationship between the training population and a new population to predict.

2.4. MATERIALS AND METHODS

2.4.1. Wheat Population

2.4.1.1. Plant Material

A total of 1,353 spring bread wheat lines from the National Agricultural Research Institute (INIA) Wheat Breeding Program of Uruguay (IWBP) were used. The INIA lines consisted of all the lines from the preliminary yield trails (F7, PYT) from 2010, 2011, and 2013; as well as the advanced (F8, AYT) and elite (F9, EYT) yield trials from 2010.

2.4.1.2. Phenotyping

Grain yield evaluation was conducted in five locations in Uruguay evaluated in four years including one location with four sowing dates. Locations used to evaluate the genotypes were Dolores (DOL, 33°50'S; 58°14'W; 15 m.a.s.l.), Durazno (DZ, 33°33' S; 56°31'W; 91 m.a.s.l.), La Estanzuela (LE, 34° 20' S; 57° 42'W; 81 m.a.s.l.), Young (YOU, 32°76'S; 57°57'W; 85 m.a.s.l.) and Ruta2 (R2, 33°45'S; 57°90'W; 95 m.a.s.l.). Four sowing dates were evaluated in LE (LE1, LE2, LE3, and LE4). The evaluation years were 2010 to 2014.

2.4.1.3. Genotyping

Genotyping by sequencing (GBS) data were obtained for the 1,353 lines from the IWBP. Tissue was collected from plants grown in either the field or greenhouse. The CTAB method (Saghai-Maroof et al., 1984) was used to isolate DNA for GBS protocol as in Poland and Rife (2012). TASSEL-GBS pipeline (Glaubitz et al., 2014) was run with a modification for non-reference genomes (Poland and Rife, 2012). Briefly, markers with a Minimum Allele Frecuency (MAF) below 1% and more than 80% of missing data were discarded. Marker-data imputation was conducted using the realized relationship matrix through the multivariate normal expectation maximization method (MVN-EM) using *rrBLUP* package (Endelman, 2011) in R software (R Development Core Team, 2015). We identified 81,999 SNPs.

2.4.1.4. Phenotypic data analysis

Phenotypic best linear unbiased estimates (BLUEs) were obtained for all genotypes present in each trial using the *nlme* package (Pinheiro et al., 2007) in R software (R Development Core Team, 2015). Field analysis was conducted according to the experimental design. Since PYT consisted on a series of smaller alpha-design trials connected through common checks, the following model was used to estimate genotypic means for each heading date group in each environment:

[1] $y_{ijkl} = \mu + \alpha_i + \lambda_j + \gamma_{k(j)} + \beta_{l(kj)} + \varepsilon_{ijkl}$

where μ is the overall mean, α_i is effect of the i-th genotype, λ_j is the effect of the j-th trial, $\gamma_{k(j)}$ is the effect of the k-replication within the j-th trial, $\beta_{l(jk)}$ is the effect of the l-th block within the j-th trial and the k-th replication, and ε_{ijkl} is the residual error from the i-th genotype in the l-th block within the k-th replication in the j-th trial; with λ_j , $\beta_{l(jk)}$ and ε_{ijkl} as random variables being $\lambda_j \sim N(0,\sigma_\lambda^2)$, $\beta_{l(jk)} \sim N(0,\sigma_\beta^2)$ and $\varepsilon_{ijkl} \sim N(0, \sigma_\varepsilon^2)$. AYT and EYT consisted on alpha-designs by heading data group and therefore, the following model was used to estimate genotypic means for each heading date group in each environment:

[2] $y_{ijk} = \mu + \alpha_i + \gamma_j + \beta_{k(j)} + \varepsilon_{ijk}$

where μ, α_i, γ_j, $\beta_{k(j)}$ and ε_{ijkl} were defined as in model [1]; with $\beta_{k(j)}$ and ε_{ijkl} as random variables $\beta_{l(jk)} \sim N(0,\sigma_\beta^2)$ and $\varepsilon_{ijkl} \sim N(0, \sigma_\varepsilon^2)$. BLUEs were estimated for each genotype in each heading date group and each environment. To avoid genotype by environment interaction, only environments belonging to Mega-Environments 1 and 2 (ME1 and ME2) identified in Lado et al. (2016) were used for this study. Mega-environments 1 and 2 include all locations of 2010, 2011, and 2013, and location LE1 of 2014. The final model to obtain genotypic means across environments was used:

[3] $y_{ijk} = \mu + \alpha_i + \delta_j + \gamma_{k(j)} + \varepsilon_{ijk}$

where μ, α_i, $\gamma_{k(j)}$, were defined as in model [1]; δ_j is j-th environment, with $\gamma_{k(j)}$ and ε_{ijk} as random variables being $\gamma_{k(j)} \sim N(0,\sigma^2_\gamma)$ and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$. BLUEs were

estimated for each genotype using using the *nlme* package (Pinheiro et al., 2007) in R software (R Development Core Team, 2015).

2.4.2 Rice Population

2.4.2.1. Plant material

A total of 644 lines from the National Agricultural Research Institute (INIA) Rice Breeding Program of Uruguay (IRBP) were used. The population consisted of 325 Indica lines, 314 Tropical Japonica lines, two Indica cultivars (*El Paso 144* and *INIA Olimar* [Blanco et al., 1993] who are the most widely grown Indica cultivars in Uruguay) and three Tropical Japonica cultivars (*INIA Parao* [Molina et al., 2011], *INIA Tacuarí* and *INIA Caraguatá* [Blanco et al, 1993]). All cultivars were used as checks in all phenotyping experiments.

2.4.2.2. Phenotyping

Rice lines were evaluated for grain yield (GY) in the Experimental Unit of Paso de la Laguna (UEPL, 33°16′S, 54°10′W), Treinta y Tres, Uruguay during the growing seasons (October-March) 2010-2011, 2011-2012 and 2012-2013.

2.4.2.3. Genotyping

Genotyping-by sequencing (GBS) data were obtained for the 644 advanced inbred lines and cultivars from the IRBP. DNA was extracted from young leaf tissue from plants grown at the Biotechnology Unit in Las Brujas, Canelones, Uruguay. The extraction was done using the Qiagen DNeasy kit (*www.qiagen.com*). GBS libraries and sequencing were done in Biotechnology Resource Center, Genomic Diversity Facility at Cornell University. Libraries were prepared using the protocol of Elshire et al. (2011). Because of the strong population structure present in lines, three data sets were obtained: an Indica, a Tropical Japonica and a combined Indica and Topical Japonica set. For both, Indica and Tropical Japonica, SNPs were called from fastq files using the TASSEL3.0 GBS pipeline (Bradbury et al., 2007) as described in Spindel et al. (2013). Alignment was performed using Bowtie 2 (Langmead and Salzberg, 2012) to the MSU version 7.0 Nipponbare rice reference genome.

Imputation of missing data was performed using the FILLIN algorithm implemented in TASSEL5.0 for Indica and Tropical Japonica genotypes separately. The average imputation accuracy was approximately 94% for both Indica and Tropical Japonica datasets. SNPs markers that had more than 50% missing data after the imputation, along with monomorphic SNPs and SNPs with a MAF below 1% were removed from the datasets. The remaining missing data were imputed by the genotypic means of each line to perform PCA analysis. For the combined Indica and Tropical Japonica set, SNPs were called from fastq files using the TASSEL3.0 GBS pipeline as described in Bradbury et al. (2007).

2.4.2.4. Phenotypic data analysis

Phenotypic best linear unbiased estimation (BLUEs) were obtained for all genotypes present in each trial using the *lme4* package (Bates et al, 2013) in R software (R Development Core Team, 2015). Field analysis was conducted according on experimental design consisting of a series of smaller randomized complete block design trials connected through common checks. The following model, using a spatial correction for row and columns was used to estimate yield genotypic means for each year and sub-species (i.e. Indica and Tropical Japonica):

$$[4] \quad y_{ijklm} = \mu + \alpha_i + \lambda_j + \beta_{k(j)} + \eta_{l(j)} + \kappa_{m(j)} + \varepsilon_{ijklm}$$

where y_{ijklm} is the response variable, μ is the overall mean, α_i is effect of the i-th genotype, λ_j is the effect of the j-th trial, $\beta_{k(j)}$ is the effect from the k-th block within the j-th trial, $\eta_{l(j)}$ is the random effect associated to the l-th row in the j-th trial, $\kappa_{l(j)}$ is the random effect associated to the l-th column in the i-th trial and ε_{ijkl}; with λ_j , $\beta_{l(jk)}$ and ε_{ijkl} as random variables being $\lambda_j \sim N(0,\sigma_\lambda^2)$, $\beta_{k(j)} \sim N(0,\sigma_\beta^2)$, $\eta_{l(j)} \sim N(0,\sigma_\eta^2)$, $\kappa_{l(j)} \sim N(0,\sigma_\kappa^2)$ and $\varepsilon_{ijklm} \sim N(0, \sigma_\varepsilon^2)$. The traits FT and GY in Tropical Japonica in 2011 were not benefited by the spatial corrections, and therefore we used model [4] without $\eta_{l(j)}$ and $\kappa_{l(j)}$ for those traits. A final model with all the environments was used to obtain overall genotypic means using the *lm* function of the *stat* package (Team, R. C., and Worldwide, C. 2002) in R software (R Development Core Team, 2015):

[5] $y_{ij} = \mu + \alpha_i + \gamma_j + \varepsilon_{ij}$

where μ is the overall mean α_i is the effect of i^{th} genotype, γ_j is the j^{th} environment and ε_{ij} as the residual error associated to each experimental unit, $\varepsilon_{ij \sim N}(0, \sigma^2_\varepsilon)$.

2.4.3. Training population optimization – within population

Different strategies were used to optimize the training population including grouping individuals by similarity or by crossing blocks. For each strategy, a G-BLUP model was fitted using rr.BLUP package (Endelman, 2011) in R software (R Development Core Team, 2015). The accuracy of the genomic estimated breeding values (GEBV) was estimated as the correlation between predicted and observed genotypic values using cross validation (Legarra et al., 2008). Models were compared in terms of accuracy to random samples of the same size, and to using of all the individuals of the training population.

General G-BLUP model: The following general model structure was used to make predictions:

[6] $y_i = \mu + \sum_{j=1}^{p} x_{ij}\, \beta_j + \varepsilon_i$

where y_i is the vector of mean yield for each genotype in all the environments (i.e. the BLUEs from a model accounting for field design and environment) of length N (N = population size or number of genotypes in the set), μ is the overall mean or intercept, x_{ij} is the allelic state of the ith individual and the jth molecular marker; βj is the effect ofthe jth molecular marker, and ε is the residual errors vector with $\varepsilon_{ij \sim N}(0, \sigma^2_\varepsilon)$. Assuming that marker effects and residual errors are independent and that $\beta_j \sim N(0, I\sigma^2_{\beta j})$, and $\varepsilon_{i \sim N}(0, I\sigma^2_\varepsilon)$ and following de los Campos and Pérez (2010) to define $u = \sum x_{ij}\beta_j$ the GBLUP model becomes:

[7] $y = 1\mu + u + \varepsilon$

where $y_{(Nx1)}$ is the vector of mean yield for each genotype in all the environments (i.e. the BLUEs from a model accounting for field design and environment) of length N (N = population size or number of genotypes in the set); $1_{(Nx1)}$ is a vector of ones of length N; μ is the overall mean; $u_{(Nx1)}$ is a random vector of genotypic predictors with

$u \sim N(0, G_{(NxN)}\, \sigma^2_g)$, where G is the realized additive relationship matrix estimated as the cross product of the centered and standardized marker states divided by the number of markers; and ε is the vector of residual errors with $\varepsilon \sim N(0, I_{(NxN)}\ \sigma^2_\varepsilon)$ where I is an identity matrix.

GBLUP model structure: We used a G-BLUP approach where marker effects could change across following Lopez-Cruz et al (2015):

[8] $$y_{ij} = \mu + \sum_{k=1}^{p} x_{ijk}\,(b_{ok} + b_{jk}) + \varepsilon_{ij}$$

where y_i is the vector of mean yield for each genotype in all the environments of length N, μ is the overall mean, x_{ij} is the allelic state of the i^{th} individual in the j^{th} molecular marker; b_{ok} is the marker effect common to all groups, b_{jk} is the random deviation of marker effect specific to the j^{th} group, and ε_{ij} is the vector of residual errors with $\varepsilon_{ij} \sim N(0,\sigma^2_\varepsilon)$. Assuming only two groups, in a matrix notation we would have:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1\mu_1 \\ 1\mu_2 \end{bmatrix} + \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} b_0 + \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

where y_1 and y_2 are vectors of mean yield for each group of length g_1 and g_2 for groups one and two respectively, $1\mu_1$ and $1\mu_2$ are vectors with the overall mean for each group with length g_1 and g_2; X_1 and X_2 matrix of molecular marker scores for each individual in each group, and ε_1 and ε_2 are the residual errors vector associated to each group; $b_o \sim N(0,I\sigma^2_{b0})$, $b_j \sim N(0,I\sigma^2_{bj})$ and $\varepsilon_i \sim N(0,I\sigma^2_\varepsilon)$.With

$$\mu = \begin{bmatrix} 1\mu_1 \\ 1\mu_2 \end{bmatrix},\ u_0 = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} b_0 \text{ and } u_1 = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

Therefore, the model can be represented as:

$$y = 1\mu + u_0 + u_1 + \varepsilon$$

where $u_0 \sim N(0,\sigma^2_{u0}\, G)$, $u_1 \sim N(0,\sigma^2_{u1}\, G_1)$ and $\varepsilon \sim N(0,I\sigma^2_\varepsilon)$, with

$$G_1 = \begin{bmatrix} \sigma^2_{u_1} X_1 X_1' & 0 \\ 0 & \sigma^2_{u_2} X_2 X_2' \end{bmatrix} / p$$

The main marker effect (u_0) allows borrowing information between groups while the group specific deviation ($u_{1)}$ captures group-specific effects. The relative importance of these two terms is determined by the corresponding variance components that are inferred from the data.

2.4.3.1 Strategy 1: Grouping based on genetic relationship

A clustering algorithm using the realized additive relationship matrix was used to group individuals in the IWBP population using the *cluster* package (Kaufman and Rousseeuw, 1990) in R software (R Development Core Team, 2015). A hierarchical agglomerative procedure using the Ward method to group similar individuals and the pseudo-F statistic to define the number of groups we used. The general G-BLUP model was used to predict breeding values within groups, and the structured G-BLUP model using groups as classes was used to predict breeding values in the whole population. Natural groups were used for the IRBP where individuals were put into one of the two sub-species: Indica or Tropical Japonica. The general G-BLUP model was used to predict breeding values within groups and a prediction for all the population regardless of group assignment. Finally, the structured G-BLUP model was also used to predict breeding values in the whole population using groups as classes.

2.4.3.2. Strategy 2: Grouping based on trials

The wheat population is evaluated in multiple trials: 2010 elite lines (F9), 2010 advanced lines (F8), 2010 preliminary lines (F7) and 2010 preliminary lines (F7). The general G-BLUP model was used to predict breeding values within groups, and the structured G-BLUP model using groups as classes was used to predict breeding values in the whole population.

2.4.3.3. Strategy 3: Grouping based on breeding objectives for the wheat population

The advanced inbred lines are routinely grouped into early or late maturity lines based on their cycle. Different breeding objectives are then pursued for each group: early maturity lines (short cycle) are selected for tighter yield based on a large

number of grains per spine, while late maturity lines (long cycle) are select for large number per spine. Therefore, yield is achieved by different yield components in the two groups.

Since different selection pressures are imposed on the two groups, the general G-BLUP model was used to predict breeding values within groups, and the structured G-BLUP model using groups as classes was used to predict breeding values in the whole population.

2.4.3.4. Cross Validation (CV)

Accuracy of genomic estimated breeding values (GEBV) was estimated as the correlation between predicted and observed genotypic values using cross validation (Legarra et al., 2008). A k-fold CV was used following Burgueño et al. (2012). Briefly, the observations were randomly divided into k non-overlapping subsets. Then, k -1 groups were used as training sets, and the remaining group was used as the validation set (i.e. GEBV are obtained for each individual in the validation set). This procedure is followed until breeding values from individuals in all k subsets are estimated. A total of 100 replications of the cross validation with k=7 were performed and the correlation between GEBV and observed genotypic values was used to estimate the accuracy of GEBV.

2.4.4. Optimization between the training population and a specific population to predict

For the optimization of the training population and for predict a specific population (TE), we used the 369 preliminary trials lines (F7 of 2013). Two strategies were chosen in the selection of training population (TR) lines taking into account only the relationship with the population to be predicted (TE).

2.4.4.1. Strategy 1: Selection of TR lines according to genetic similarity matrix (G) with population to be predicted (TE)

Line selection was made to integrate the candidate training population based on the genetic relationship matrix (G) between TR and TE. Matrix G was calculated from

the matrix of molecular markers and the choice of the lines was made according to three criteria: first, the TR lines were ordered according to the mean of their genetic relationship with the TE. 15% of the TR lines with the highest average TE, 20% and so on were added by adding the lines of five in five (25%, 30%, ...) until reaching 100% of the TR. Second, the TR lines were ordered according to the medians of the genetic relationship with the TE and the previous procedure was repeated for the choice of lines. Finally, all lines that had a genetic relationship value greater than or equal to the first quartile of the empirical distribution of the maximum values of relationship between populations were chosen. Each of these population subgroups for the three criteria were used to train the G-BLUP model and predict TE.

2.4.4.2. Strategy 2: Selection of TR lines according to genetic similarity matrix weighted by the estimation of molecular marker effects (Gw)

The genetic distance of the TR lines with the TE lines was calculated as follows (Endelman, 2011):

$$SD_{ij} = \sqrt{\sum_{k=1}^{k=M} \left(\frac{\left| X_{\cdot PE[i,k]} - X_{\cdot PP[j,k]} \right|}{2} \right) * \hat{u}[k]^2}$$

where i is the i-th line of the TR, j is the j-th line of the TE, and k is the k-th marker; $X_{\cdot PE(n.PE*m)}$ y $X_{\cdot PP(n.PE*m)}$ are the genotypic matrices with the state of each marker (-1, 0, 1); M the number of markers and u is the vector of the estimation of the effects of the molecular markers obtained from a mixed model with the *mixed.solve* function of the rrBLUP library (Endelman, 2011) in R (Development Core Team, 2008).

First, for each line of the training population, we calculated the mean distance with the population to be predicted weighted by the estimated effect of the markers. Lines were sorted from lowest to highest according to that average distance value and 15% of highest relationship was selected, 20%, 25% ..., etc., from 5% in 5% to 100% of the lines of the TR. Each of these population subgroups was used to train the G-

BLUP model and predict TE. Second, the same procedure was performed, but with the median distances of each of the TR lines with the population to be predicted. Third, all lines that had a genetic distance value less than or equal to the third quartile of the empirical distribution of the minimum values of genetic distance between populations were chosen.

Finally, a random selection of TR lines was performed to predict TE, taking subgroups of the same size as those obtained according to the two strategies. Subgroup with 15% of the lines, 20% and jumping from 5% in 5% to 100%, G-BLUP models were trained for each subset and predicted to TE.

2.5. RESULTS

2.5.1. Training population optimization - within population

Wheat population

The greater the number of lines in the training population, the greater the accuracy of the predictions (Fig. 1). From 600 lines the increase in precision becomes marginal. For the wheat population and according to the three clustering strategies, some groups presented higher accuracy than those obtained with random samples of the same size or the entire training population (Fig. 1). These groups are: group 1 (G1), obtained from cluster analysis based on the genetic relationship matrix, AYT, the final trials of the 2010 program and the short cycle lines (SC).

For the population of wheat, the model that includes the structure of the groupings showed greater accuracies when modeling the groups of lines coming from the different traits and when modeling the groups of lines according to cycle. We obtained an accuracy of 0.488 for the first and 0.444 for the second. Both accuracies are greater than that obtained with the model that does not include the structure and that used all the lines of the training population (Fig. 1).

Figure 2, 3 and 4 show the genetic structure and the description of the most frequent parent names for the three grouping strategies, as well as the origin of the lines in some cases. A greater genetic structure is observed Inside group 1 (Fig. 3) than in the remaining groups resulting from the cluster analysis. This group shows greater surface of red color outside the diagonal representing the greater genetic distance in the diagonal groups of lines that are most similar to each other. A similar but less marked behavior can be seen in the graphs of the final and short cycle lines. On the other hand, there is agreement between the families of the lines of group 1 and those of the short cycle lines.

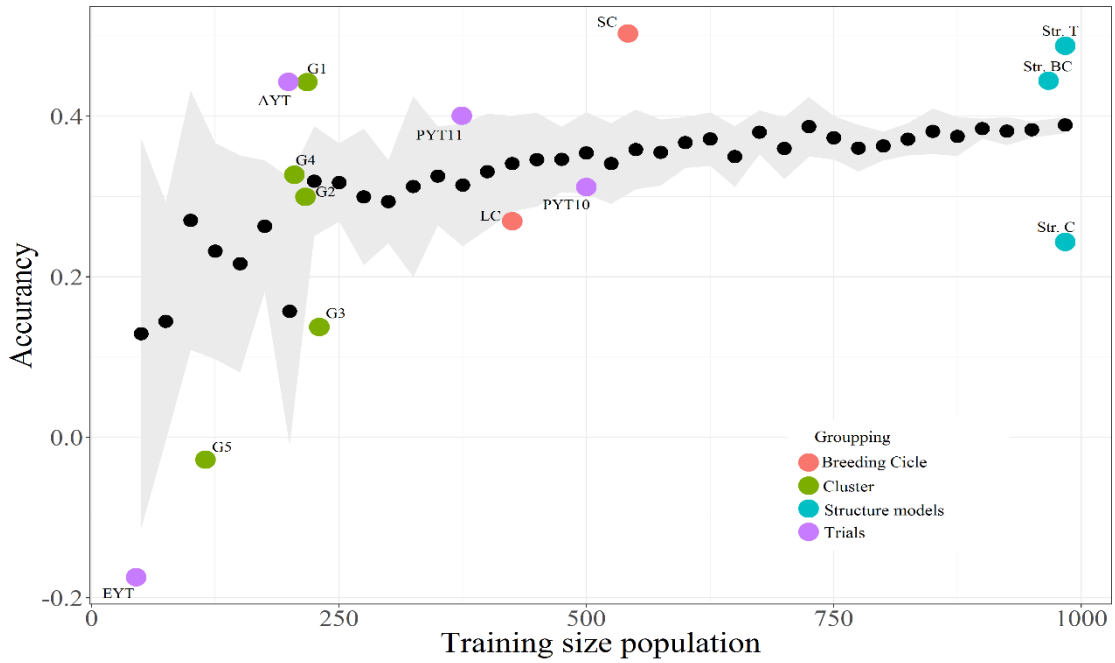


Figure 1: Accuracy of the predictions of the three clustering strategies used in the optimization of Training Population within the Wheat Population.

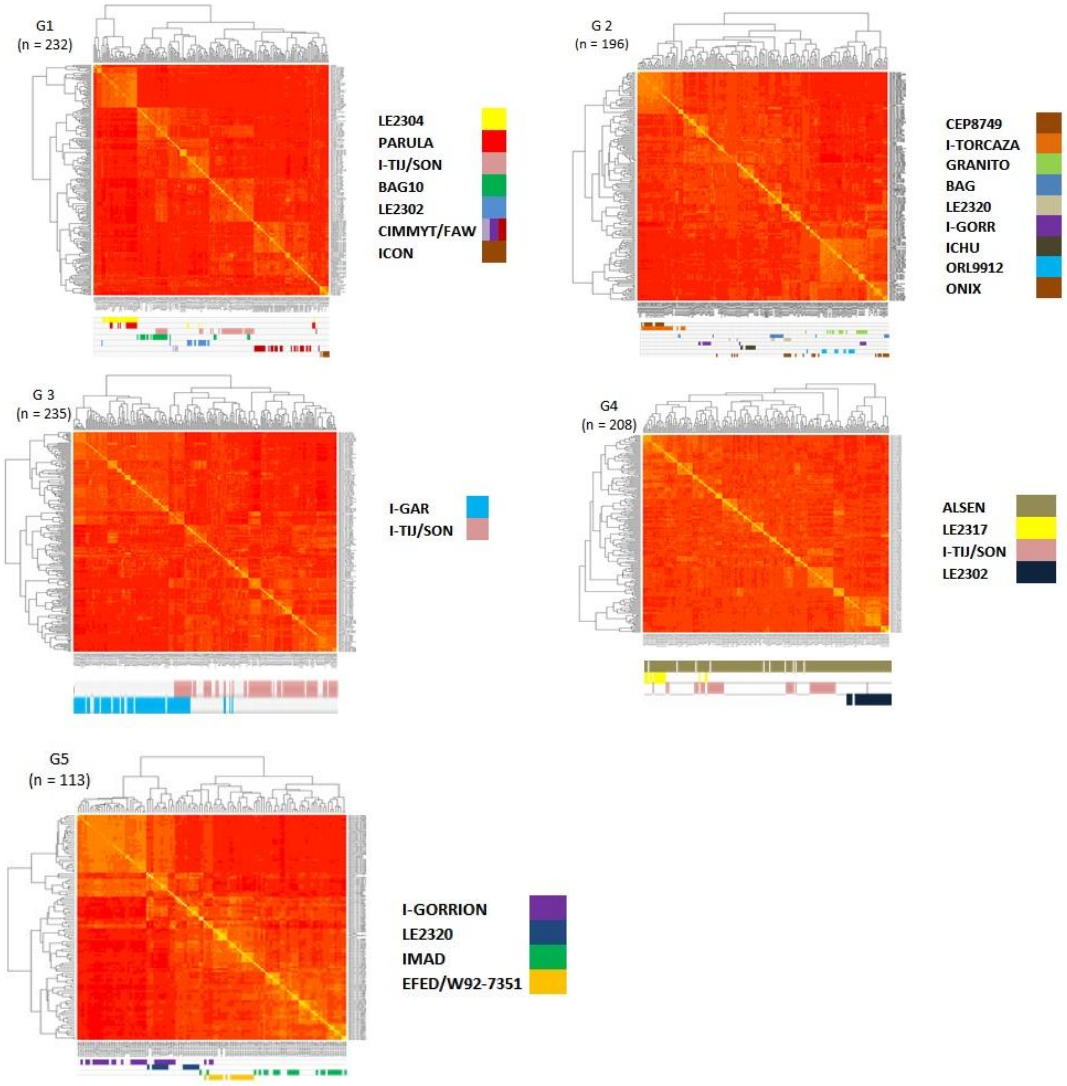


Figure 2: Heatmap of the genetic relationship matrix and representation of families within each of the groups generated by the cluster analysis for the wheat population.

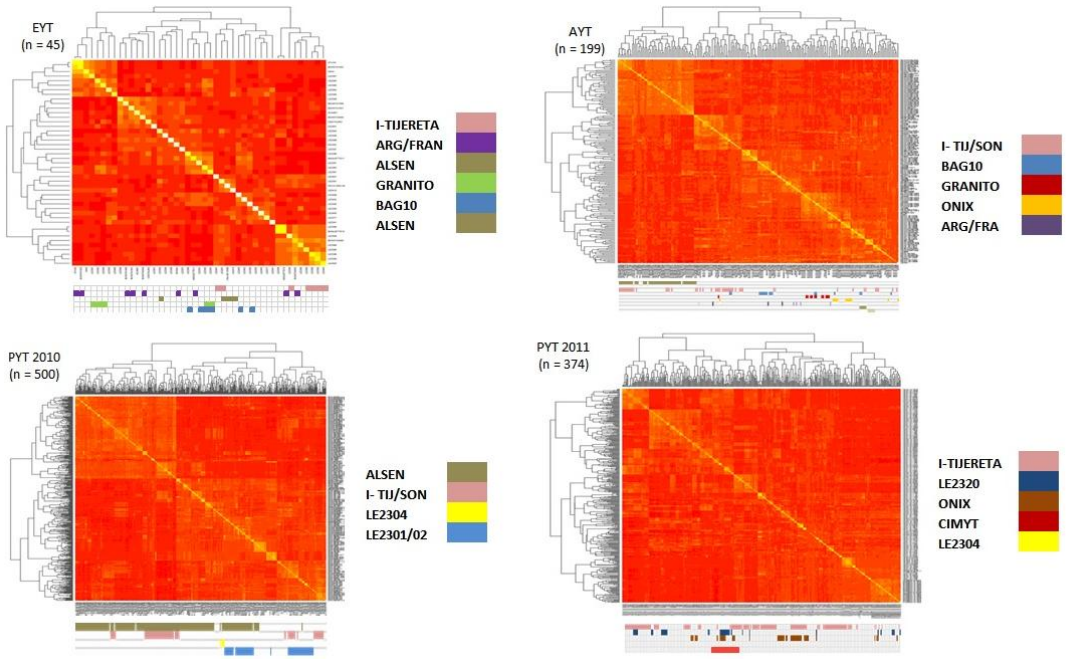


Figure 3: Heatmap of the genetic relationship matrix and representation of families for each of the trials for the wheat population.

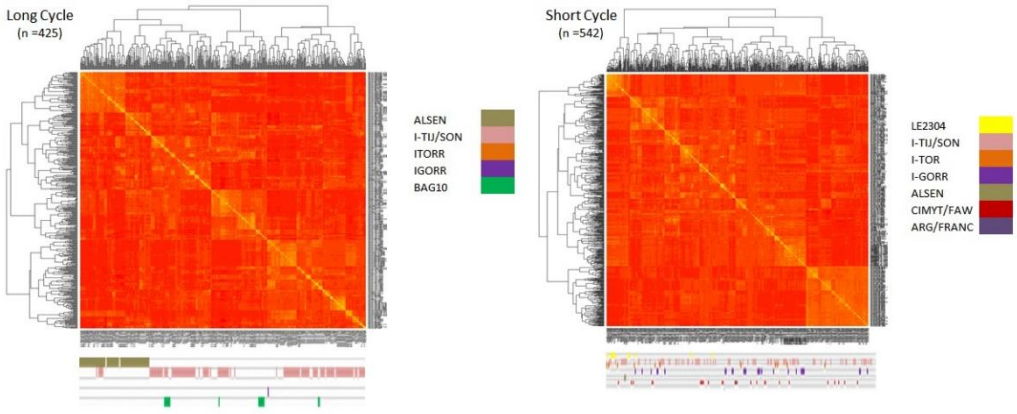


Figure 4: Heatmap of the genetic relationship matrix and representation of families for the lines grouped by cycle.

Rice population

For the Rice population, the model that considers the 2 subspecies together showed accuracies of 0.8 for all the considered population sizes. The model that includes the structure is not different from the one that does not include it (Fig. 5). By modeling within each subspecies using genotypic information separately, the accuracies of the predictions resulted in the environment of 0.5 (Fig. 5).

The heatmap of the Rice population shows the strong structure in subspecies and its strong association with the phenotypic character (grain yield). This aspect of the population is also reflected in figure 7, which presents a diagram of the performance observed vs the predicted performance for the joint model.

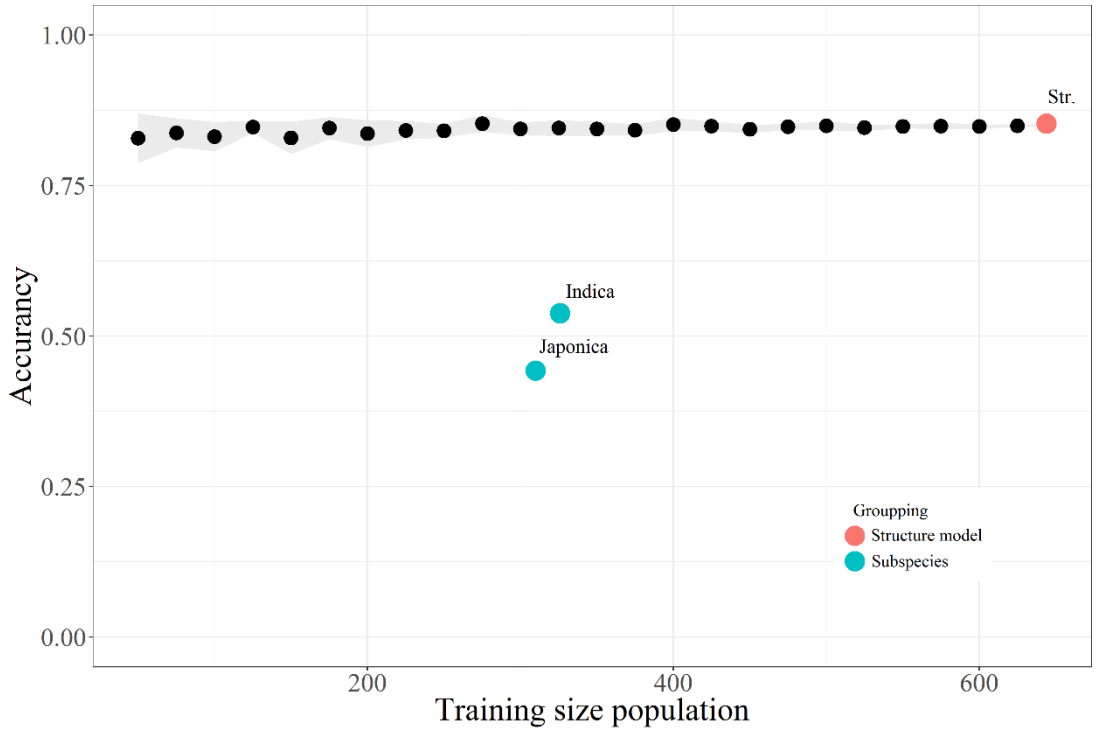


Figure 5: Precision of the predictions within the training population of Rice with the groupings according to subspecies.

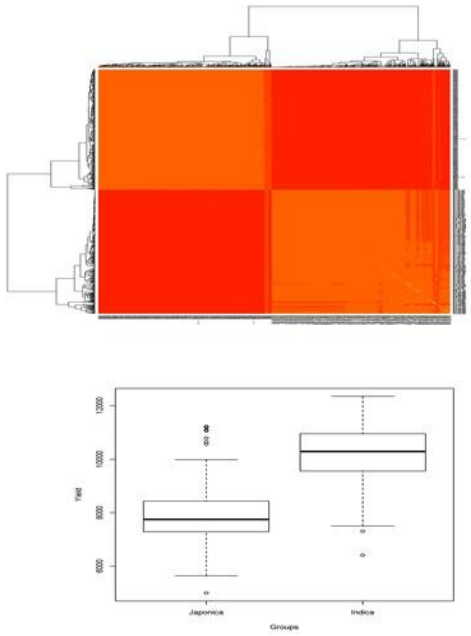


Figure 6: Heatmap of genetic relationship matrix and yield by subspecies.

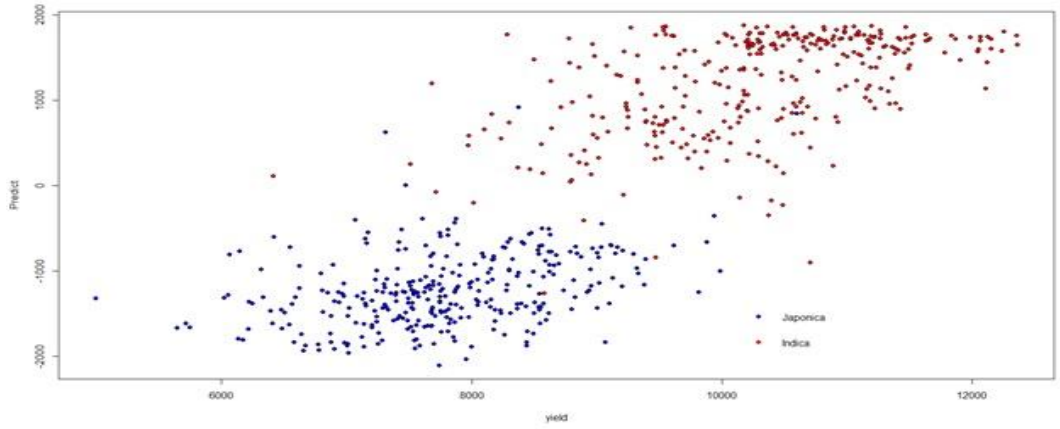


Figure 7: Observed and predictor performance of the model for yield by subspecies.

2.5.2. Optimization between the training population and a given population to predict specific

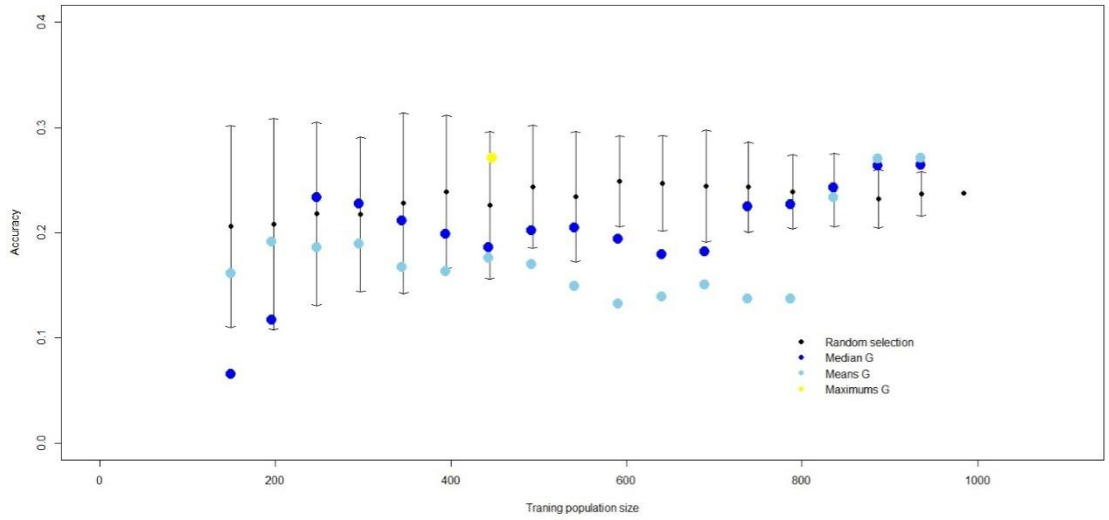
To predict genetic values for a new population in wheat was found that using the estimation of the effect of the markers to calculate the genetic distance between the lines of TR and TE improves the accuracy of the predictions (Figure 6). Better precision accuracies were obtained for all subsets of TR when this strategy was used (Fig. 6).

Several subsets of TR lines were found that showed greater accuracies in TE predictions than that obtained when using the entire training population. These subgroups consist of 25%, 30% of the TR lines most related to TE and in 70% of the lines according to the criterion that estimates the effect of the markers (Fig. 6).

For our small sizes, the average of the distances between the lines weighted by the effect of the markers optimizes the accuracy of the model. We also obtain a greater accuracy in the predictions for large samples with the median of the weighted distances (Fig. 5 and Fig. 6).

The accuracy of the estimated model with the TR composed of lines that present the maximum genetic relationship with the population to be predicted proved to be greater than that obtained with the entire training population (Fig. 8).

(a)



(b)

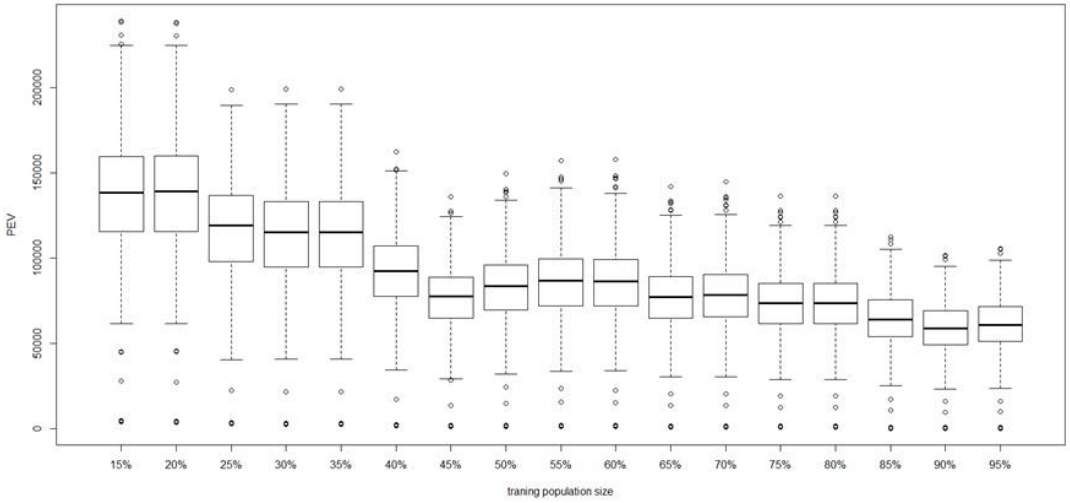
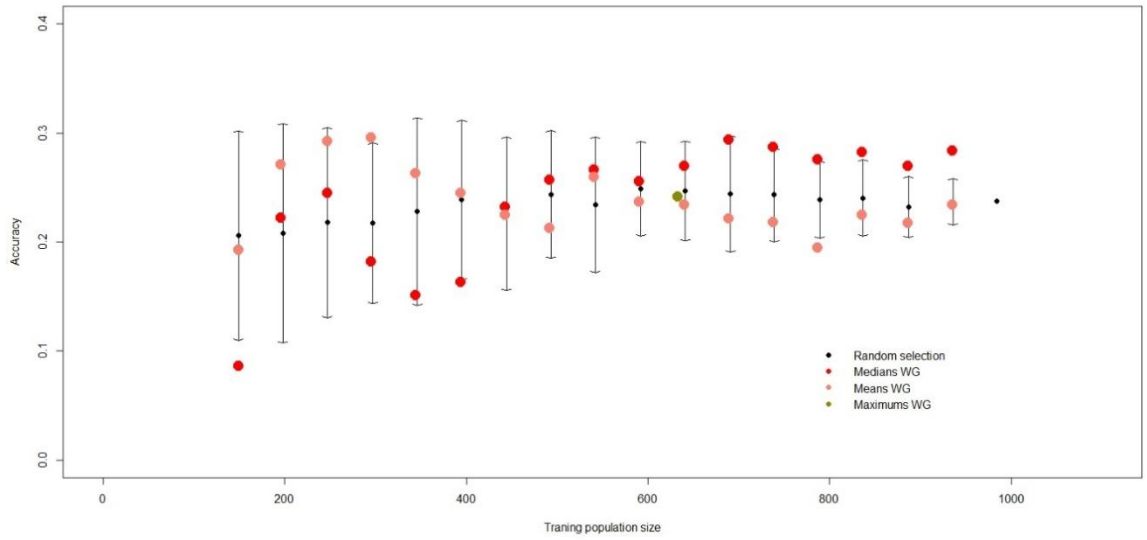


Figure 8: (a) The accuracy of the predictions for distance calculation strategy based on genetic relationship matrix and (b) prediction error variances (PEV) for different sizes of training population according to the mean criteria of choice of population lines of training.

(a)



(b)

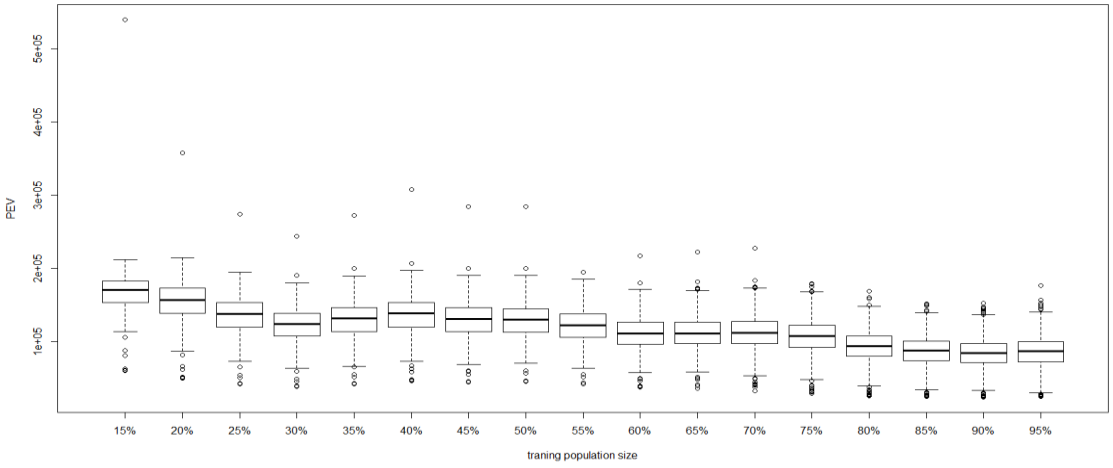


Figure 9: (a) The accuracy of the predictions for the distance calculation strategy based on the molecular markers weighted by the estimation its effects on their own and (b) prediction error variances (PEV) for the different sizes of the training population according to the mean criteria of choice of the lines of the training population.

2.6. DISCUSSION

The training population can be optimized in terms of size (Meuwissen et al., 2009; Asoro et al., 2011; Isidro et al., 2015), relationship among individuals (Habier et al, 2007; Asoro et al., 2011; Clark et al., 2012; Isidro et al., 2015) and population structure (Asoro et al., 2011; Crossa et al., 2014). We found an increase in predicition accuracy with larger population sizes. This has been shown in other studies as well (Lorenzana and Bernardo, 2009, Asoro et al., 2011). On the other hand, it has been widely documented that training populations that are more related to the prediction populations show higher prediction accuracies for a given population size (Clark et al, 2012; Lorenz et al, 2012; Habier et al, 2007). However, some studies that showed that increasing the population size even at the expense of lower genetic relationships might provide higher prediction accuracies (Asoro et al., 2011). Furthermore, Crossa et al (2014) indicated that groups of individuals with higher average relationship matrix showed higher precision accuracy in cross-validation. However, we found that evaluating the genetic relationship with the realized relationship matrix (G matrix) did not provide a good idea of the prediction ability of the population. Groups of individuals with larger values of genetic relationship on average were not consistently better predicted than groups with smaller values on average. On the other hand, the groups of individuals with the highest prediction accuracy were not the ones with the highest average relationship matrix. The best prediction accuracy in our study was found for groups with some population structure (i.e. G1, FYT, and SC) in the form of families. G1 (n=232) based on genetic clustering of individuals is formed by seven families or closely related individuals: advanced lines derived from crosses to CV. INIA-Tijereta, INIA-LE2304 advanced line, CV. Parula, as well as CIMMYT and French germplasm. FYT (n=199) is formed by five families or closely related individuals: advanced lines derived from crosses to CV. INIA-Tijereta, Granito, Onix or Baguette10 cultivars, and Argentinean and French germplasm. On the other hand, FYT is a more diverse group because it also contains some common checks, and several sister lines. Finally, the SC (n=542) group is also composed of seven families or closely related

individuals: advanced lines derived from crosses to CV. INIA-Tijereta, INIA_LE2304 advanced line, CV. INIA-Torcaza, CV. ALSEN, and French, Argentinean and CIMMYT germplasm. We believe that having groups of related individuals in the training population provides a closely related individual for every individual in the population to predict. This contradicts some findings by Riedelsheimer et al., (2013) that showed that predicting within families was better than having several families.

Toosi et al. (2010) showed that prediction accuracy was higher for individuals of the same cattle race, but showed only a small decrease in prediction accuracy for multi-race training populations, indicating that population structure can be accounted for if the LD structure is similar (Asoro et al., 2011). These models work well as long as allelic effects are predictive from one population to the next one (Lorenz et al., 2011). Combining sub-groups would therefore increase population size (Asoro et al., 2011) and might compensate the slight decrease in prediction accuracy. Finally, families with INIA-Tijereta and INIA-LE2304 offspring, as well as French, Argentinean, and CIMMYT germplasm were present in the high prediction accuracy groups and were therefore individuals with more predictive ability than others. We believe that in our study, the G matrix did not appropriately capture the pedigree relationships and therefore was not a good predictor of prediction accuracy. Pedigree information should be evaluated in GS models (Crossa et al., 2010).

Modeling population structure can be successfully accomplished in GS as long as the same LD relationship structures are maintained across populations (Asoro et al., 2011; Toosi et al., 2010; Lopez-Cruz et al., 2015). We found better prediction accuracies when modeling population structure including either trial or cycle information in our models.

However, population structure can be a problem when markers are in opposite phases among sub-populations as in the case of rice. We found a strong population structure, associated to the phenotype that artificially inflated the prediction accuracy (i.e. r=0.85) by being very accurate in predicting group belonging, but being very bad in predicting within groups (i.e. r<0.5). The over-estimation of prediction accuracy was

also observed by Schmidt et al. (2016) under evidence the population structure caused by two subgroups in barley population.

Our first objective was to see if we could identify a set of groups that would be more related and will therefore overcome the effect of population size by having higher prediction accuracy due to their increased relatedness. We were only partially successful at this because we found some groups that had higher prediction accuracy, but the results were not consistent for a given partitioning. We believe that part of the reason for increasing prediction accuracy in some groups was the presence of family structure that guaranteed a closely related individual in the TP for every individual in the TE. However, these results were obtained by using classical CV1 and training and predicting within a single population. Therefore, to truly mimic the breeding program structure, we decided to evaluate strategies to select individuals as TP for a given TE. The average genetic relationship matrix (G) was not a good indicator for prediction accuracy; however, the weighted average genetic relationship matrix (Gw) was a better strategy.

Several studies have documented that adding unrelated lines to the TP decreased the prediction accuracy (Lorenz et al., 2015). On the other hand, in some cases, the genetic relationship could be somewhat offset by larger population sizes (Clark et al, 2012;). The increase in population size could offset the genetic relationship when historic LD is acounted for, and population sizes and marker density is high (Habier et al., 2013; Hickey et al., 2015; Meuwissen, 2009). Therefore, a larger mixed population with high marker density would be a better option than a smaller more related population (Asoro, 2011). Marker density is important to increase the probability of finding markers in LD with a given QTL across populations (Daetwyler et al., 2010). We found a group of 246 individuals that produced the optimal prediction accuracy. This group was identified by including the 25 to 30% of the individuals from the TP that had the highest mean weighted genetic relationship to the TE. Additionally, after 700 individuals, high prediction accuracies are obtained regardless of the method used to include those individuals. Larger population sizes also produced smaller prediction error variances.

2.7. REFERENCES

Asoro, F. G., Newell, M. a., Beavis, W. D., Scott M. P., Jannink, J.L. 2011. Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. The Plant Genome Journal 4: 132-144.

Bates, D., Sarkar, D., 2010 Linear mixed-effects models using S4 classes. R Package *lme4*: http://ftp.auckland.ac.nz/software/CRAN/doc/packages/lme4.pdf.

Blanco, P., Pérez de Vida, F., Piriz, M. 1993. Inia-Tacuarí Nueva variedad de arroz precoz de alto rendimiento. INIA (Instituto Nacional de Investigación Agropecuaria). Boletín de Divulgación N° 31.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T. M., Ramdoss, Y., Buckler, E. S. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.

Burgueño, J., de los Campos, G., Weigel, K., Crossa, J. 2012. Genomic Prediction of 505 Breeding Values when Modeling Genotype × Environment Interaction using Pedigree 506 and Dense Molecular Markers. Crop Science. 52:707-719.

Clark, S.A., Hickey, J.M., Daetwyle, H.D., van der Werf, J.H.J. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genetic Selection Evolution 44:1-9.

Combs, E., Bernardo, R. 2013. Accuracy of genome wide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6:1-7.

Crossa, J., de los Campos, G., Pérez, P., Gianola D., Burgueño, J. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–24.

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112: 48–60.

Daetwyler, H. D, Pong-Wong, R., Villanueva, B., Woolliams, J. A. 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185: 1021-1031.

de los Campos, G., Pérez, P. 2010. BGLR: Bayesian Generalized Linear Regression R Package: http://cran.rproject.org/web/packages/BGLR/BGLR.pdf.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., Calus, M. 2012. Whole genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327–345.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Plos One 6: 1–10.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 4:250– 255.

Glaubitz, J.C., Casstevens, T. M., Lu F., Harriman, J., Elshire, R.J., Sun, Q, Buckler, E. S. 2014. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. Plos One 9:1–11.

Habier, D., Fernando, R. L., Dekkers, J. C. M. 2007. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177: 2389–2397

Habier, D., Fernando, R.L., Garrick, D.J. 2013. Genomic BLUP decoded: A look into the black box of genomic prediction. Genetics 194:597–607.

Heffner, E.L., Jannink, J.L., and Sorrells, M.E. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4:65–75.

Heslot, N., Yang, H., Sorrells, M.E., and Jannink, J.L. 2012. Genomic selection in plant breeding: A comparison of models. Crop Science. 52:146–160.

Hickey, J., Dreisigacker, M. S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., Grondona, M., Zambelli, A., Windhausen, V. S., Mathews, K., Gorjanc, G. 2015. Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. Crop Science 54:1476-1488.

Isidro, J., Jannink, J.L., Akdemir, D., Poland, J, Heslot, N., Sorrells, M.E. 2015. Training set optimization under population structure in genomic selection Theoretical and Applied Genetics 128:145–158.

Kaufman, L., Rousseeuw, P.J. 1990. Cluster analysis methods. R Package 'cluster' Version 2.0.1. http://cran.r-project.org/web/packages/cluster/cluster.pdf.

Langmead, B., Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359.

Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen. 2008. Performance of genomic 621 selection in mice. Genetics 180: 611–618.

Lopez-Cruz, M, Crossa, J., Bonnett, D., Dreisigacker, S., Poland J., Jannink, J, Singh , R.P, Enrique Autrique, E, de los Campos, G. 2015. Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker x Environment Interaction Genomic Selection Model. G3 5: 569–582.

Lorenz, A.J., Chao, S., Asoro, F., Heffner, E., Hayashi, T., Iwata, H., Smith, K., Sorrels, M., Jannink, J.L. 2011. Genomic selection in plant breeding: Knowledge and prospects. En: Sparks D.L. (ed.). Advances in agronomy. San Diego: Academic Press. 77–123.

Lorenz, A. J., Smith, K. P., Jannink, J.L. 2012. Potential and Optimization of Genomic Selection for Fusarium Head Blight Resistance in Six-Row Barley. Crop Science, 52: 1609-1621.

Lorenz, A.J. 2013. Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: A simulation experiment. G3: Genes, Genomes, Genet. 3:481–491.

Lorenz, A.J., Smith, K. P. 2015. Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. Crop Science 55: 2657-2667.

Lorenzana, R.E., Bernardo, R. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theoretical and Applied Genetics 120:151–161.

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

Meuwissen ,T.H. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genetics Selection Evolution, 41:35.

Molina, F., Blanco, P., Pérez de Vida, F., 2011. Nuevo cultivar de arroz INIA L5502 PARAO: características y comportamiento. INIA (Instituto Nacional de Investigación Agropecuaria). Arroz. 68:26-32.

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Animal Breeding and Genetics 124: 342-355.

Pinheiro, J., Bates, D., Roy, S. D., Sarkar, D. 2007. Linear and Nonlinear Mixed Effects Models. R Package *nlme.* ftp://ftp.uni-bayreuth.de/pub/math/statlib/R/CRAN/doc/packages/nlme.pdf

Poland, J.A., Rife, T.W. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome 5: 92–102.

Pszczola, M., Strabel, T., Mulder, H.A, Calus, M.P.L. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. Dairy Science. 95:389–400.

R Development Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Riedelsheimer, C., Endelman, J.B., Stange, M., Sorrells, M.E., Jannink, J., Melchinger, A.E. 2013. Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503.

Saghai-Maroof, M. A., Soliman, K .M., R A Jorgensen, Allard, R.W. 1984. Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proceedings of the National Academy of Science of the United State of America 24:8014–8018.

Schmidt, M., Kollers, S., Maasberg-Prelle, A., Grober, J., Schinkel, B., Tomerius, A., Graner, A., Korzun, V. 2016. Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. Theoretical and Applied Genetics 129: 203–213.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., Meuwissen, T.H. 2008. Genomic selection using different marker types and densities. Animal Science 86:2447–2454.

Spindel, J., Wright, M., Chen, C., Cobb, J., Gage, J., Harrington, S., Lorieux, M., Ahmadi, N., McCouch, S. 2013. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theoretical and Applied Genetics 126: 2699–2716.

Team, R. C., and Worldwide, C. 2002. The R stats package. R Foundation for Statistical Computing, Vienna, Austria: Disponible en: http://www. R-project. org.

Toosi, A., Fernando, R.L., Dekkers, J.C.M. 2010. Genomic selection in admixed and crossbred populations. Animal Science. 88:32–46.

Wientjes, Y.C.J., Veerkamp, R.F., Calus, M.P.L. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:621–631.

Xu, S. 2003. Estimating polygenic effects using markers of the entire genome. Genetics, 163: 789-801.

3. RESULTADOS Y DISCUSIÓN GENERAL

La población de entrenamiento puede ser optimizada en términos del tamaño (Meuwissen et al., 2009; Asoro et al., 2011; Isidro et al., 2015), de la relación entre los individuos (Habier et al, 2007; Asoro et al., 2011; Clark et al., 2012; Isidro et al., 2015) y estructura de la población (Asoro et al., 2011; Crossa et al., 2014). Nuestros resultados muestran que con mayores tamaños de la población de entrenamiento la exactitud de las predicciones se incrementa. Esto también ha sido mostrado en otros estudios (Lorenzana y Bernardo, 2009, Asoro et al., 2011).

Si bien se ha documentado ampliamente que poblaciones de entrenamiento más relacionadas con poblaciones a predecir muestran mayores precisiones en la predicción para un tamaño de población dado (Clark et al., 2012; Habier et al, 2007), algunos estudios muestran que el aumento del tamaño de la población, incluso con menores relaciones genéticas, podría incrementar la exactitud en las predicciones (Asoro et al., 2011). Además, Crossa et al. (2014) indicó que grupos de individuos con promedios altos en la matriz de relacionamiento mostraron mayor exactitud de precisión en la validación cruzada. Sin embargo, nosotros encontramos que medir el relacionamiento genético con la matriz G no resultó ser una buena estrategia para elegir las mejores líneas para entrenar los modelos de selección genómica.

Los grupos de líneas con mayores valores promedios de relación genética no fueron consistentemente mejores predictores que los grupos con valores promedios menores. Por otro lado, los grupos de líneas con mayor precisión de predicción no fueron los que tenían mayor promedio de relacionamiento genético. La mejor precisión de predicción en nuestro estudio se encontró para grupos con alguna estructura de población en forma de familias. Estos grupos son: (1) grupo 1 (n = 232) obtenido a partir del análisis de cluster basado en la matriz de relacionamiento genético, está formado por cinco familias o grupos de líneas estrechamente relacionadas: líneas avanzadas derivadas de cruzamientos con el cultivar INIA-Tijereta, línea avanzada INIA-LE2304, cultivar Parula, así como CIMMYT y germoplasma francés (Fig. 2), (2) las líneas del grupo de ensayos finales del programa del año 2010 FYT (n = 199) formado por cinco familias o líneas

estrechamente relacionadas: líneas avanzadas derivadas de cruzamientos con los cultivares INIA-Tijereta, Granito, Onix o Baguette10 y germoplasma argentino y francés. Por otro lado, FYT es un grupo más diverso porque también contiene algunos testigos comunes, y varias líneas hermanas (Fig. 3) y (3) el grupo de líneas de ciclo corto SC (n = 542) compuesto también de siete familias o individuos estrechamente relacionados: líneas avanzadas derivadas de cruzamientos con los cultivares INIA-Tijereta, línea avanzada INIA_LE2304, cultivar INIA-Torcaza, cultivar ALSEN, y germoplasma francés, argentino y de CIMMYT (Fig.4).

Al tener varios grupos más relacionados dentro de la población de entrenamiento, con la validación cruzada, en el momento de la predicción siempre habrá líneas estrechamente relacionadas con la población a predecir. Esto contradice algunos hallazgos de Riedelsheimer et al. (2013) que demostraron que la predicción dentro de las familias era mejor que tener varias familias.

Toosi et al. (2010) mostraron que predecir líneas dentro de una misma raza de ganado resulta en altas precisiones, sin embargo, al predecir con poblaciones de entrenamiento multi-razas las precisiones tuvieron una leve disminución. Por lo tanto, si la estructura de desequilibrio de ligamiento (LD) es similar a través de las razas, la estructura de la población compuesta puede ser modelada con los modelos de selección genómica (Asoro et al., 2011). Estos modelos funcionan bien siempre y cuando los efectos alélicos sean predictivos de una población a la siguiente (Lorenz et al., 2011). Así mismo, la combinación de subgrupos aumentaría el tamaño de la población (Asoro et al., 2011) y compensando la ligera disminución en la precisión de la predicción. En nuestros resultados las familias con descendientes de INIA-Tijereta y INIA-LE2304, así como germoplasma francés, argentino y de CIMMYT estuvieron presentes en los grupos con mejor performance en la predicción y por lo tanto son líneas con mayor capacidad de predicción que otras. En nuestro estudio, la matriz G no resultó ser un buen predictor de la exactitud de las predicciones ya que no captura apropiadamente las relaciones de pedigrí. Incluir la información de pedigrí en los modelos de selección genómica GS podría ser beneficioso (Crossa et al., 2010).

Modelar la estructura de la población puede mejorar las predicciones en selección genómica, siempre que se mantengan las mismas estructuras de relación LD entre las poblaciones (Asoro et al., 2011; López-Cruz et al., 2015). Encontramos mayor precisión de predicción al modelar la estructura de la población, cuando incluimos la información de ensayo y de ciclo en nuestros modelos. Se obtuvo una precisión de 0,488 para el primero y 0,444 para el segundo. Ambas precisiones son mayores que la obtenida con el modelo que no incluye la estructura y que utiliza todas las líneas de la población de entrenamiento (Fig. 1). El rendimiento en granos está claramente asociado tanto al ciclo como al ensayo.

Sin embargo, la estructura de la población puede ser un problema cuando los marcadores están en fases opuestas entre las subpoblaciones como en el caso del arroz. La fuerte estructura de la población asociada al fenotipo infló artificialmente la precisión de la predicción (r = 0,85), siendo muy precisa la predicción de pertenencia a la subespecie, pero siendo mala en predecir dentro de la subespecie (r <0,5) (Fig. 7). La sobreestimación de la exactitud de predicción también fue observada por Schmidt et al. (2016) en cebada.

Nuestro primer objetivo fue ver si podíamos identificar un conjunto de grupos que estarían más relacionados y, por lo tanto, superar el efecto del tamaño de la población por tener mayor exactitud de predicción debido a su mayor relacionamiento genético. La estructura familiar presente en los grupos fue en parte la razón de la buena performance en las predicciones. La estructura familiar garantiza que exista un individuo estrechamente relacionado en la TP para cada individuo en la TE. Sin embargo, estos resultados se obtuvieron utilizando la validación cruzada clásica dentro de la población. Por lo tanto, para imitar verdaderamente la estructura del programa de mejoramiento, decidimos evaluar dos estrategias para seleccionar las líneas de la población de entrenamiento y predecir una nueva población. El promedio de la matriz de relación genética (G) no fue un buen indicador para la exactitud de la predicción, sin embargo, el promedio de la matriz de relación genética ponderada por el efecto estimado de los marcadores (Gw) fue una mejor estrategia.

Varios estudios han documentado que la adición de líneas no relacionadas con el TP disminuyó la precisión de la predicción (Lorenz et al., 2015). Por otro lado, en

algunos casos, la relación genética podría ser algo compensada por mayores tamaños de población (Clark et al, 2012). El aumento del tamaño de la población podría contrarrestar la relación genética cuando se considera la LD histórica, el tamaño de la población y una alta densidad de marcadores moleculares (Meuwissen, 2009; Habier et al., 2013). Por lo tanto, una población mixta más grande con alta densidad de marcadores sería una mejor opción que una población más pequeña y más relacionada (Asoro, 2011). La densidad de marcadores es importante para aumentar la probabilidad de encontrar marcadores en LD con el QTL entre poblaciones (Daetwyler et al., 2010). Encontramos un grupo de 246 líneas con las que se obtuvo la mayor precisión en la predicción. Este grupo está compuesto por el 30% de las líneas de la población de entrenamiento más cercanas a la población a predecir cuándo medimos la cercanía como el promedio de la matriz de relacionamiento ponderada. Además, a partir del 70% de las líneas más cercanas, se obtienen altas precisiones de predicción independientemente del método utilizado para incluir a esos individuos. La exactitud del modelo estimado con la PE compuesta por líneas que presentan el máximo relacionamiento genético con la población a predecir resultó ser mayor a la obtenida con toda la población de entrenamiento (Fig. 8). Finalmente, para las dos estrategias la varianza del error de predicción promedio resultó ser menor al incrementar el tamaño poblacional. Para la población de trigo, mayor tamaño poblacional optimiza las predicciones. Los tamaños de población más grandes también produjeron variaciones de error de predicción más pequeñas.

5. BIBLIOGRAFÍA

Asoro F. G., Newell M. a., Beavis W. D., Scott M. P., Jannink J.L. 2011. Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. The Plant Genome Journal 4: 132.

Bates D., Sarkar D., 2013. Linear mixed-effects models using S4 classes. R Package *lme4*: http://ftp.auckland.ac.nz/software/CRAN/doc/packages/lme4.pdf.

Bernardo R., Yu J. 2007. Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Science 47:1082-1090.

Bernardo R. 2008. Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. Crop Science 48: 1649-1664.

Blanco P., Pérez de Vida F., Piriz M. 1993. Inia-Tacuarí Nueva variedad de arroz precoz de alto rendimiento. INIA (Instituto Nacional de Investigación Agropecuaria). Boletín de Divulgación N° 31.

Bradbury P.J., Z. Zhang Z., Kroon D.E., Casstevens T. M., Ramdoss Y., Buckler E. S. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa. 2012. Genomic Prediction of 505 Breeding Values when Modeling Genotype × Environment Interaction using Pedigree 506 and Dense Molecular Markers. Crop Sci. 52: 707-719.

Clark S.A., Hickey J.M., Daetwyle H.D., van der Werf J.H.J 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genetic Selection Evolution 44:4.

Combs E., and Bernardo R. 2013. Accuracy of genome wide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6:1-7.

Crossa J., de los Campos G., Pérez P., Gianola D., Burgueño J. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–24.

Crossa J., Pérez P., Hickey J., Burgueño J., Ornella L. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112: 48–60.

Daetwyler H. D, Pong-Wong R., Villanueva B., Woolliams J. A. 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185: 1021-1031.

de los Campos G. and Pérez P. 2010. BGLR: Bayesian Generalized Linear Regression R Package: http://cran.rproject.org/web/packages/BGLR/BGLR.pdf.

de los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D., Calus M. 2012. Whole genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327–345.

de Roos A.P., Hayes B.J., Spelman R.J., Goddard M.E. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179: 1503-1512.

DIEA-MGAP (Estadísticas Agropecuarias, Ministerio de Ganadería Agricultura y Pesca), 2014. Anuario 2014. Disponible en: http://www.mgap.gub.uy/portal/page.aspx?2,diea,diea-anuario-2014,O,es,0,

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Plos One 6: 1–10.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 4:250–255.

FAO (Food and Agriculture Organization of the United Nations), 2014. Food and Nutrition in Numbers 2014. Disponible en: http://www.fao.org/3/a-i4175e.pdf

Glaubitz J.C., Casstevens T. M., Lu F., Harriman J., Elshire R.J., Sun Q, Buckler E. S. 2014. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLOS ONE 9:1–11.

Gianola D., de los Campos G., Hill W.G., Manfredi E., Fernando R.L. 2011. Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363.

Habier D., Fernando R. L., Dekkers J. C. M. 2007. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177: 2389–2397.

Habier D., Fernando R.L., Garrick D.J. 2013. Genomic BLUP decoded: A look into the black box of genomic prediction. Genetics 194:597–607.

Hawkesford M.J., Araus J.L., Park R., Calderini D., Miralles D. 2013. Prospects of doubling global wheat yields. Food and Energy Security 2: 34–48.

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

Heffner E.L., Jannink J.L., and Sorrells M.E. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4:65–75.

Heslot N., Yang H., Sorrells M.E., and Jannink J.L. 2012. Genomic selection in plant breeding: A comparison of models. Crop Science. 52:146–160.

Hickey J., Dreisigacker M. S., Crossa J., Hearne S., Babu R., Prasanna B. M., Grondona M., Zambelli A., Windhausen V. S., Mathews K., Gorjanc G. 2015. Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. Crop Science 54:1476-1488.

Isidro J., Jannink J.L., Akdemir D., Poland J, Heslot N., Sorrells M.E. 2015. Training set optimization under population structure in genomic selection Theoretical and Applied Genetics 128:145–158.

Jannink J.L., Lorenz A.J, Iwata H. 2010. Genomic selection in plant breeding: From theory to practice. Briefing Functional Genomics 9:166–177.

Kaufman L. and Rousseeuw P.J. 1990. Cluster analysis methods. R Package ‘cluster’ Version 2.0.1. http://cran.r-project.org/web/packages/cluster/cluster.pdf.

Lado B., González P., Quincke M., Silva P., Gutiérrez L., 2016. Modeling Genotype × Environment Interaction for Genomic Selection with Unbalanced Data from a Wheat Breeding Program. Crop Science 56:2165-2179.

Lande R. and Thompson R. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743-756.

Langmead B. and Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359.

Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen. 2008. Performance of genomic 621 selection in mice. Genetics 180(1): 611–618.

Lopez-Cruz M, Crossa J., Bonnett D., Dreisigacker S., Poland J., Jannink J, Singh R.P, Enrique Autrique E, de los Campos G. 2015. Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker x Environment Interaction Genomic Selection Model. G3: Genes, Genomes, Genet. 5: 569–582.

Lorenz A.J., Chao S., Asoro F., Heffner E., Hayashi T., Iwata H., Smith K., Sorrels M., Jannink J.L. 2011. Genomic selection in plant breeding: Knowledge and prospects. En: Sparks D.L. (ed.). Advances in agronomy. San Diego: Academic Press. 77–123.

Lorenz A. J., Smith K. P. Jannink J.L. 2012. Potential and Optimization of Genomic Selection for Fusarium Head Blight Resistance in Six-Row Barley. Crop Science, 52: 1609-1621.

Lorenz A.J. 2013. Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: A simulation experiment. G3: Genes, Genomes, Genet. 3:481–491.

Lorenz A.J. and Smith K. P. 2015. Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. Crop Science 55: 2657-2667.

Lorenzana R.E. and Bernardo R. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theoretical and Applied Genetics 120:151–161.

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

Meuwissen T.H. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genetics Selection Evolution, 41:35.

Molina F., Blanco P., Pérez de Vida F. 2011. Nuevo cultivar de arroz INIA L5502 PARAO: características y comportamiento. INIA (Instituto Nacional de Investigación Agropecuaria). Arroz. 68:26-32.

Muir W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Animal Breeding and Genetics 124: 342-355.

Pinheiro J., Bates D., Roy S. D., Sarkar D. 2007. Linear and Nonlinear Mixed Effects Models. R Package *nlme.* ftp://ftp.uni-bayreuth.de/pub/math/statlib/R/CRAN/doc/packages/nlme.pdf

Poland J.A. and Rife T.W. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome 5: 92–102.

Pszczola M., Strabel T., Mulder H.A, Calus M.P.L. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. Dairy Science. 95:389–400.

R Development Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Riedelsheimer C., Endelman J.B., Stange M., Sorrells M.E., Jannink J., Melchinger A.E. 2013. Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503.

Saghai-Maroof M. A., Soliman K .M., R A Jorgensen, Allard R.W. 1984. Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proceedings of the National Academy of Science of the United State of America 24:8014–8018.

Schmidt M., Kollers, S., Maasberg-Prelle A., Grober J., Schinkel , B., Tomerius A., Graner A., Korzun V. 2016. Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. . Theoretical and Applied Genetics 129: 203–213.

Solberg T. R., Sonesson A. K., Woolliams J. A., Meuwissen T.H.E. 2008. Genomic selection using different marker types and densities. Animal Science 86:2447-2454.

Spindel J., Wright M., Chen C., Cobb J., Gage J., Harrington S., Lorieux M., Ahmadi N., McCouch S. 2013. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theoretical and Applied Genetics 126: 2699–2716.

Team, R. C., and Worldwide, C. 2002. The R stats package. R Foundation for Statistical Computing, Vienna, Austria: Disponible en: http://www. R-project. org.

Toosi A., Fernando R.L., Dekkers J.C.M. 2010. Genomic selection in admixed and crossbred populations. Animal Science. 88:32–46.

Wientjes Y.C.J., Veerkamp R.F. and Calus M.P.L. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:621–631.

Xu S. 2003. Estimating polygenic effects using markers of the entire genome. Genetics, 163: 789-801.