

**UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE AGRONOMÍA**

**ESTRATEGIAS PARA LA APLICACIÓN DE SELECCIÓN
GENÓMICA EN UN PROGRAMA DE MEJORAMIENTO DE TRIGO**

por

Bettina LADO LINDNER

TESIS presentada como uno de los
requisitos para obtener el título de
Doctora en Ciencias Agrarias

**MONTEVIDEO
URUGUAY
julio 2018**

Tesis aprobada por el tribunal integrado por Ing. Agr. Ph.D. Pablo Speranza, Ing. Agr. Ph.D. Silvia Germán y Ing. Agr. Ph.D. Francisco Peñagaricano, el (día) de (mes) de (año). Autora: B.C. Mag. Bettina Lado. Director/a Ing. Agr. Ph.D. Lucía Gutiérrez, Co-director/a Ph.D. Jesse Poland y Ing. Agr. Ph.D. Martín Quincke.

AGRADECIMIENTOS

Los agradecimientos finales de un trabajo de tesis son difíciles de realizar, aunque es una linda experiencia pasar nuevamente por tantos lugares y personas. En este acto se cierra un proceso muy largo que ha involucrado muchas personas y aspectos de mi vida. Como temo dejar mucha gente importante por fuera he decidido no nombrar personas específicas.

No hay duda de que esta tesis nunca podría haber sido hecha tal cual está sin la enorme colaboración de mis tutores y comité de seguimiento, principalmente mi directora académica quien es parte de este trabajo.

Agradezco a las instituciones que directa e indirectamente financiaron esta tesis. A la Universidad de la República que me permitió dedicar horas de mi trabajo al doctorado, así como la posibilidad y el tiempo de realizar pasantías en otras Universidades y asistir a congresos de difusión de este trabajo. Comisión Sectorial de Investigación Científica y sus programas de movilidad que financiaron estos viajes e intercambios fundamentales en el extranjero. Agencia Nacional de Investigación e Innovación quien financió la beca de este doctorado. Instituto Nacional de Investigación Agropecuaria quien financió el proyecto de tesis y gracias a muchos compañeros y amigos de esta institución quienes formaron parte de este proceso.

Agradezco a Kansas State University y University of Wisconsin-Madison por permitirme realizar pasantías, a sus investigadores y a muchos compañeros que conocí allí y de quienes aprendí mucho.

Gracias a mis compañeros y amigos del Departamento de Biometría, Estadística y Cómputos a los que están hoy y los que se han ido. Gracias por la compañía y el apoyo en lo académico y lo emocional, por ayudarme a hacer de las múltiples piedras en este camino perlas que son la gran cosecha de este Doctorado.

Gracias amigos y familia, por el apoyo incondicional y por ser parte de esto, porque sin ustedes esta tesis no tendría sentido.

TABLA DE CONTENIDO

PÁGINA DE APROBACIÓN.....	II
AGRADECIMIENTOS.....	III
RESUMEN.....	VI
SUMMARY.....	VII
1. <u>INTRODUCCIÓN</u>	1
1.1. HIPÓTESIS DE INVESTIGACIÓN.....	6
1.2. OBJETIVOS.....	8
1.2.1. <u>Objetivo general</u>	8
1.2.2. <u>Objetivos específicos</u>	8
1.3. ESQUEMA GENERAL DE TESIS.....	8
2. <u>MARCO TEÓRICO</u>	10
2.1. MEJORAMIENTO GENÉTICO VEGETAL.....	10
2.1.1. <u>Mejoramiento genético de plantas autóгамas</u>	11
2.2. SELECCIÓN ASISTIDA POR MARCADORES.....	17
2.2.1. <u>Marcadores Moleculares</u>	17
2.2.2. <u>Mapeo de QTL</u>	23
2.3. SELECCIÓN GENÓMICA.....	25
2.3.1. <u>Modelos de predicción</u>	26
2.3.2. <u>Población de entrenamiento</u>	31
2.4. INTERACCIÓN GENOTIPO POR AMBIENTE.....	33
2.5. SELECCIÓN POR CARACTERES MÚLTIPLES.....	37
2.5.1. <u>Correlación entre caracteres</u>	37
2.5.2. <u>Respuesta correlacionada a la selección</u>	38
2.5.3. <u>Métodos de selección por caracteres múltiples</u>	38
2.5.4. <u>Predicciones de caracteres múltiples en SG</u>	40
2.6. PREDICCIÓN DE CRUZAMIENTOS.....	41
3. <u>MODELADO DE LA INTERACCIÓN GENOTIPO POR AMBIENTE</u>	43
3.1. RESUMEN DEL CAPÍTULO.....	43
4. <u>OPTIMIZACIÓN DE RECURSOS EN LA PREDICCIÓN DE MÚLTIPLES CARACTERES DE CALIDAD PANADERA DE TRIGO</u>	59

4.1. RESUMEN DEL CAPÍTULO.....	59
4.2. INTRODUCTION.....	63
4.3. MATERIALS AND METHODS.....	66
4.3.1. <u>Plant Material</u>	66
4.3.2. <u>Phenotyping</u>	67
4.3.3. <u>Phenotypic Analyses</u>	67
4.3.4. <u>Genotyping</u>	68
4.3.5. <u>Genomic prediction models</u>	69
4.3.5.1 Single-trait model.....	69
4.3.5.2 Multi-trait models.....	70
4.3.6. <u>Cross-validation scheme</u>	70
4.3.7. <u>Improving efficiency of phenotyping</u>	71
4.4. RESULTS.....	73
4.4.1. <u>Phenotypic characterization of the population</u>	73
4.4.2. <u>Multi-trait genomic predictions</u>	75
4.4.3. <u>Multi-trait predictions using correlated traits</u>	75
4.4.4. <u>Replace phenotyping</u>	76
4.5. DISCUSSION.....	79
4.5.1. <u>Multi-trait genomic predictions</u>	79
4.5.2. <u>Predictions for replace phenotyping</u>	80
4.6. CONCLUSION.....	82
4.7. REFERENCES.....	83
5. <u>COMPARACIÓN DE ESTRATEGIAS DE PREDICCIÓN DE CRUZAMIENTOS PARA LA SELECCIÓN DE LOS MEJORES CRUZAMIENTOS EN UN PROGRAMA DE MEJORAMIENTO GENÉTICO DE TRIGO</u>	89
5.1. RESUMEN DEL CAPÍTULO.....	89
6. <u>DISCUSIÓN GENERAL</u>	103
7. <u>CONSIDERACIONES FINALES Y PERSPECTIVAS</u>	109
8. <u>BIBLIOGRAFÍA</u>	111

RESUMEN

La inclusión de selección genómica (SG) en un programa de mejoramiento permite acortar los tiempos, incrementando la ganancia genética por unidad de tiempo. SG puede ser usado en la predicción de los mejores cruzamientos evaluando la complementariedad alélica entre los padres y/o prediciendo el desempeño de individuos. Los modelos mixtos utilizados en SG son una herramienta poderosa que permite incluir correlaciones entre individuos, ambientes y caracteres permitiendo utilizar toda la información disponible. El objetivo de este trabajo es optimizar herramientas para la inclusión de SG en el programa de mejoramiento genético de trigo (PMGT), para el manejo de la interacción genotipo por ambiente (IGA), la predicción de múltiples caracteres y la selección de los mejores cruzamientos. En este trabajo se utilizaron datos históricos del PMGT de Uruguay, proveniente de 35 ambientes, 1495 líneas experimentales evaluadas para rendimiento y 695 líneas evaluadas para caracteres de calidad panadera de trigo. Se compararon modelos multi-carácter y multi-ambiente modelando correlaciones en la predicción de nuevos individuos e individuos parcialmente fenotipados. También se predijeron los mejores cruzamientos considerando y no considerando la varianza genética entre ambos padres. Los resultados muestran que los modelos multi-ambiente y el multi-carácter son útiles para la optimización de recursos de un programa de mejoramiento, prediciendo líneas parcialmente fenotipadas. Finalmente, seleccionar cruzamientos considerando la media del desempeño de los padres fue tan bueno como considerar además la varianza genética entre ambos padres para rendimiento de grano; no fue así en los cruzamientos seleccionados para calidad panadera dónde hubo diferencias entre ambos métodos.

Palabras clave: Interacción genotipo por ambiente, caracteres múltiples, selección de cruzamientos, valor genético, modelos mixtos.

STRATEGIES TO APPLY GENOMIC SELECTION INTO A WHEAT BREEDING PROGRAM

SUMMARY

The inclusion of genomic selection (GS) in a breeding program allows us to shorten breeding time, increasing the genetic gain per unit of time. GS can be used to predict the best crosses, accounting for allelic complementation between parents and/or predicting individuals' performance. Mixed models used in GS are a powerful tool to include individuals' correlations, environments and traits, allowing the use of the whole available information. The goal of this study is the optimization of GS as a tool in the wheat breeding program, to manage the genotype by environment interaction, multi-trait prediction, and best crosses selection. Historical data from the national wheat breeding program were used; data were obtained from 35 environments, 1495 experimental lines, evaluated for yield; and 695 lines evaluated for baking quality. Multi-trait and multi-environment models were compared modeling correlations to predict new individuals, and individuals partially phenotyped. Best crosses were also predicted considering or not considering the variance between parents. Results showed that multi-environment and multi-trait models were useful to optimize the breeding program resources, by predicting partially phenotyped lines. Finally, crosses selection considering parent's mean performance was as good as including the variance in selection for grain yield, but not for quality traits.

Keywords: genotype by environment interaction, multiple traits, crosses selection, genetic value, mixed models.

1. INTRODUCCIÓN

El trigo es un cultivo de importancia mundial en términos económicos y productivos dado su rol principal en la alimentación humana. El área cultivada de trigo es la mayor en el mundo en comparación con los principales cultivos (arroz, soja, maíz y cebada), con entre 212-220 millones de hectáreas cosechadas por año en los últimos 10 años (FAO 2017). Aproximadamente el 60% del total de la producción de trigo es usada para alimentación humana (Bonjean et al. 2016), en la cual el trigo contribuye con aproximadamente el 18% del suministro alimentario de energía (kcal per cápita por día) y el 20% de las proteínas (g per cápita por día), aportando también fibras y vitaminas (FAO, 2017). En Uruguay, el trigo ha sido históricamente el cultivo de mayor importancia y actualmente ocupa el segundo lugar en superficie sembrada, después de la soja (DIEA 2017).

El mejoramiento genético en distintas especies cultivadas ha sido la estrategia por la cual se han obtenido cultivares con mejores rendimientos, buena respuesta a enfermedades y con calidad industrial, que a su vez son estables dentro de determinadas regiones del mundo (Duvick 2005; Trethowan et al. 2005; Graybosch y Peterson 2010; Hawkesford et al. 2013). Un ejemplo de mejoramiento genético de plantas exitoso fue el desarrollo de cultivares semi-enanos de trigo, insensibles al fotoperiodo y con resistencia a distintas enfermedades, lo que permitió obtener variedades con altos rendimientos y amplia adaptación (Trethowan et al. 2007). El proceso de obtención de una nueva variedad en el mejoramiento de plantas autóгамas como el trigo, se inicia con cruzamientos entre líneas promisorias con el objetivo de generar diversidad. Luego, la progenie se autofecunda por varias generaciones con el objetivo de obtener líneas homogéneas, homocigotas para la mayoría de sus alelos (líneas puras) para que sean estables luego de múltiples generaciones de producción de semilla (Baenziger y Depauw 2009). Durante el proceso de obtención de líneas puras las plantas pueden o no ser evaluadas fenotípicamente para su selección por características de interés. El proceso de obtención de una nueva variedad de trigo a partir de un cruzamiento inicial oscila entre 10-12 años para el mejoramiento genético tradicional, lo cual puede reducirse a 5-6 años utilizando dos siembras por año en distintas localidades (también llamado siembra fuera de estación o *shuttle breeding*;

Ortiz et al. 2007). Uno de los desafíos en el mejoramiento de trigo es acortar los tiempos de obtención de nuevas variedades adaptadas (Hesey et al. 2002).

Dentro de las estrategias disponibles para la obtención de líneas puras, los métodos de descendencia de semilla única (DSU) y haploides duplicados (HD) permiten avanzar generaciones rápidamente y obtener líneas homocigotas en menor tiempo que los métodos tradicionales de mejoramiento (ej. bulk, masal, pedigrí) (Brim 1966; Laurie y O'Donoghue 1994; Guzy-Wrobelska y Szarejko 2003). Una de las limitantes de estas metodologías es que al aplicarse a gran escala, necesitan una gran cantidad de recursos. Otra limitante es la imposibilidad de realizar una evaluación fenotípica hasta que se obtienen líneas puras, debido a la insuficiente disponibilidad de semillas (Baenziger y Depauw 2009).

Los marcadores moleculares se han desarrollado en el contexto del mejoramiento de plantas y son especialmente útiles para seleccionar individuos cuando no es posible realizar una evaluación fenotípica (Lande y Thompson 1990). Los marcadores por los que se selecciona se encuentran ligados (se heredan juntos) a las regiones del genoma que están involucradas en la expresión del carácter de interés. Por lo tanto, al seleccionar individuos que poseen los marcadores moleculares de interés se selecciona indirectamente los individuos que poseen las regiones génicas favorables para dicho carácter. La identificación de los marcadores ligados a las regiones del genoma involucradas en la expresión de caracteres cuali y cuantitativos requiere pasos previos de análisis de ligamiento (para caracteres cualitativos), análisis de locus de carácter cuantitativo (QTL) o mapeo asociativo (para caracteres cuantitativos). Estos análisis permiten identificar si el marcador está o no ligado a algún gen o QTL involucrados en la expresión del carácter de interés. Por otro lado, la selección genómica (SG) es una estrategia de selección para caracteres cuantitativos que utiliza marcadores moleculares, sin realizar una identificación previa de los marcadores ligados a los QTL, dado que analiza el efecto de miles de marcadores a lo largo de todo el genoma. La SG predice el desempeño de nuevas líneas experimentales utilizando su información genotípica y un modelo estadístico “entrenado” (Meuwissen et al. 2001). Este modelo estadístico integra información fenotípica y genotípica de un grupo de líneas conocidas, denominada población de entrenamiento.

La SG presenta ventajas en la selección de líneas para caracteres fenotípicos complejos, especialmente cuando la información fenotípica: 1) no es confiable debido a la baja heredabilidad de esos caracteres (ej. evaluación fuera de época); 2) no está disponible (ej. poca semilla); o 3) requiere de metodologías más complejas o más recursos para su determinación en comparación con el genotipado (ej. determinaciones de calidad industrial). Por lo tanto, se espera que la aplicación de la SG, en combinación con los métodos de rápido avance generacional permita acelerar la tasa de ganancia genética y acortar los ciclos de mejoramiento (Heffner et al. 2010; Bassi et al. 2016). Estudios empíricos y de simulación muestran que la SG, tiene una adecuada precisión en la estimación de los valores genéticos de los individuos para distintos cultivos como trigo (Heffner et al. 2011; Poland et al. 2012; Michel et al. 2016; Hayes et al. 2017), maíz (Crossa et al. 2013; Lyra et al. 2017), arroz (Spindel et al. 2016; Wang et al. 2017) y otras especies (Resende et al. 2011; Kumar et al. 2012b; Zhang et al. 2016).

Una de las etapas más importantes del mejoramiento genético es el cruzamiento entre dos individuos promisorios, etapa en la cual se genera la diversidad genética en la progenie a partir de la cual se realizará la selección (Baenziger y Depauw 2009). La selección de las mejores cruzas es una etapa clave del mejoramiento en la cual se dispone de poca información y muchos individuos para elegir (Bernardo 2003). La mejor craza no siempre es entre los dos individuos con mejor desempeño, ya que el desempeño de la progenie dependerá de cómo se complementen genéticamente ambos padres (Souza y Sorrells 1991). O sea que si dos padres presentan buen desempeño y sus alelos son complementarios su descendencia tendrá un mayor potencial de acumular alelos favorables para el carácter de interés que dos padres con alelos idénticos. Una medida de la complementariedad alélica es la varianza genética entre ambos padres, la cual puede calcularse considerando que los marcadores están en equilibrio de ligamiento (se asume que los marcadores se heredan de forma independiente entre sí) o considerando el desequilibrio de ligamiento presente en la población (Endelman 2011; Mohammadi et al. 2015). La SG tiene un gran potencial de aplicación en la selección de estos cruzamientos, ya que utiliza modelos de predicción genómica para establecer la mejor combinación de padres (Endelman 2011;

Bernardo 2014; Mohammadi et al. 2015; Tiede et al. 2015). Sin embargo, la predicción de los mejores cruzamientos es una herramienta potencial dentro del mejoramiento genético, cuya utilización aún no ha sido validada.

Para la aplicación de SG en un programa de mejoramiento es necesario predecir adecuadamente el desempeño de nuevos individuos (sin información fenotípica para ningún carácter ni ambiente) reuniendo la información disponible de varios años, localidades y caracteres (Cooper et al. 2014). Para la predicción de caracteres múltiples mediante simulación, se ha demostrado que, en la predicción de individuos nuevos, los modelos que integran la información de varios caracteres son más adecuados que aquellos que predicen un carácter a la vez (Jia y Jannink 2012; Guo et al. 2014). Sin embargo, esta ventaja de los modelos multi-carácter aún no se ha corroborado utilizando datos experimentales en plantas (Jia y Jannink 2012; Dos Santos et al. 2016; Schulthess et al. 2016; Sun et al. 2017). Por otro lado, la inclusión de datos de distintos ambientes requiere evaluar la presencia de interacción genotipo por ambiente (IGA), esto quiere decir que el *ranking* o las magnitudes de la diferencias del desempeño de los genotipos pueden ser distintas en diferentes ambientes. Por lo tanto, al integrar la información multi-año y multi-localidad para predecir nuevos individuos, es necesario considerar la presencia de IGA. Burgueño et al. (2012) fueron los primeros en desarrollar modelos que incorporan la IGA en el contexto de SG, siguiendo la misma estrategia de los modelos multi-carácter. Para definir la mejor estrategia de predicción de nuevas líneas es necesario estudiar la IGA característica de cada región o programa de mejoramiento (Windhausen et al. 2012; Dawson et al. 2013). A continuación en el proceso de selección de nuevos individuos es necesario comparar entre estrategias que ignoren la IGA (prediciendo medias fenotípicas en el total de los ambientes) o la consideren (prediciendo dentro de mega-ambientes o utilizando modelos multi-ambiente) en la predicción de nuevos individuos (Burgueño et al. 2012; Dawson et al. 2013).

Los modelos de predicción multi-carácter y multi-ambiente son útiles para predecir también el comportamiento de individuos en ambientes desconocidos o para caracteres para los cuales no han sido evaluados, dado que incorporan las varianzas y covarianzas entre individuos, caracteres y/o ambientes (Burgueño et al. 2012; Guo

et al. 2014). Este tipo de predicción de individuos parcialmente evaluados puede ser una ventaja para optimizar los recursos de fenotipado y genotipado dentro de un programa de mejoramiento (Guo et al. 2014). Se ha demostrado la ventaja de los modelos multi-carácter en la predicción de caracteres de alto costo o difíciles de medir, utilizando caracteres correlacionados de bajo costo y simple medición (Guo et al. 2014; Rutkoski et al. 2012; Sun et al. 2017). Debido a esto, estos modelos pueden ser útiles en la predicción de caracteres complejos como calidad panadera en trigo, la cual está determinada por varios caracteres correlacionados entre sí y con diferentes costos de medición (Battenfield et al. 2016). Sin embargo, para utilizar estos modelos y optimizar recursos, sería necesario definir una estrategia de predicción que permita fenotipar menos individuos y utilizar la información genómica para remplazar el fenotipado aplicando las predicciones. Los modelos multi-ambiente también ayudan a optimizar recursos ya que permiten fenotipar más líneas en menos ambientes y predecir su valor genético en ambientes desconocidos (Lorenz 2013; Endelman et al. 2014). Burgueño et al. (2012) demostraron que, cuando existe información de individuos en ambientes correlacionados, incorporar las covarianzas ambientales en los modelos de SG, mejora la precisión de las predicciones del valor genético en determinado ambiente. Para utilizar esta estrategia de optimización de recursos son necesarios más estudios que permitan determinar cuál es la mejor forma de predecir entre grupos de ambientes.

En este trabajo se buscó optimizar las herramientas para la aplicación de SG en un programa de mejoramiento genético de trigo, para lo que se evaluó si la precisión de las predicciones de nuevos individuos mejora cuando se incluye la IGA en los modelos de predicción. Los modelos que incluyen o no IGA se compararon en presencia de distintos niveles de IGA (grupos de predicción por localidades, años y mega-ambientes) y con distintos tamaños de población de entrenamiento. A continuación se evaluaron los modelos multi-caracter y multi-ambiente en la optimización de la asignación de recursos de un programa de mejoramiento. Para optimizar los recursos de fenotipado en múltiples-ambientes se comparó la habilidad predictiva de modelos que incluyen e ignoran la IGA en la predicción de individuos en ambientes desconocidos (Capítulo 3). En el contexto de caracteres múltiples, se

evaluó la optimización del fenotipado de un carácter costoso y de difícil medición, comparando modelos multi-carácter con distintos niveles de fenotipado de dicho carácter. En el mismo sentido, se optimizó el fenotipado de los caracteres correlacionados más económicos y de fácil medición (Capítulo 4). Finalmente, se determinó el rol de la SG en la selección de los mejores cruzamientos. Para esto se comparó la predicción de cruzamientos utilizando el valor genético medio de los padres (como predictor de la media de la progenie) y la media del 10% superior de la progenie, la cual considera la varianza genética de la progenie. Se comparó la selección de cruza utilizando dos formas de predicción de la varianza: considerando la población en equilibrio o desequilibrio de ligamiento (Capítulo 5).

1.1. HIPÓTESIS DE INVESTIGACIÓN

Capítulo 3: Modelado de la IGA utilizando datos desbalanceados para la aplicación de SG en un programa de mejoramiento de trigo

La predicción del valor genético de nuevos individuos o de individuos fenotipados en ambientes distintos de los que se quieren predecir se puede realizar ignorando, reduciendo (a través del uso de mega-ambientes) o incluyendo la IGA en los modelos de predicción. Nuestras hipótesis son que:

- Modelos que incluyen la IGA predicen mejor a nuevos individuos (individuos no fenotipados) que modelos sin IGA (la precisión de las predicciones de modelos con IGA es superior a la precisión de las predicciones de modelos sin IGA).
- Modelar la IGA no ayuda a predecir nuevos individuos cuando se emplea dentro de mega-ambientes que reducen la IGA (la precisión de las predicciones de los modelos con y sin IGA es similar dentro de mega-ambientes).
- Modelar la IGA siempre es mejor para predecir individuos parcialmente fenotipados (la precisión de las predicciones de modelos con IGA es superior a la precisión de las predicciones de modelos sin IGA cuando existen individuos fenotipados en otros ambientes).

Capítulo 4: Optimización de la asignación de recursos utilizando predicciones genómicas en caracteres múltiples de calidad panadera en trigo.

Mediante el uso de los modelos multi-carácter es posible incrementar la información disponible prediciendo de forma conjunta caracteres correlacionados entre sí. Nuestras hipótesis son que:

- Modelos multi-carácter pueden ser utilizados para reemplazar el fenotipado y predecir un carácter costoso o de compleja determinación con menos recursos utilizando información de caracteres correlacionados más económicos o de fácil medición.
- Los modelos multi-carácter mantienen su habilidad predictiva al disminuir considerablemente (ej. 50%) los individuos fenotipados para caracteres correlacionados.
- Utilizando la misma cantidad de datos fenotípicos es más eficiente fenotipar más individuos desbalanceadamente (fenotipar diferentes individuos para distintos caracteres) que fenotipar balanceadamente menos individuos (fenotipar los mismos individuos para todos los caracteres).

Capítulo 5: Estrategias para la selección de cruzamientos utilizando predicciones genómicas en el mejoramiento de trigo.

Es posible seleccionar los mejores cruzamientos considerando únicamente el desempeño de los padres, que es un buen predictor del desempeño medio de la progenie, o considerando también la varianza genética entre ambos padres. A su vez, la varianza genética puede predecirse considerando los marcadores en equilibrio o desequilibrio de ligamiento. Nuestras hipótesis son que:

- Seleccionar padres considerando la predicción de la varianza del cruzamiento permitirá obtener mayores ganancias genéticas en la siguiente generación en comparación con seleccionar padres en función de su valor de cría *per se*.
- Existen diferencias en la selección de los mejores cruzamientos si se predice la varianza de las cruza tomando en cuenta el desequilibrio de ligamiento o considerando que la población está en equilibrio de ligamiento.

- Al aumentar la influencia de la varianza en la selección de las mejores cruzas es posible seleccionar líneas más diversas sin reducir la ganancia genética en la siguiente generación.

1.2. OBJETIVOS

1.2.1. Objetivo general:

Optimizar herramientas para la aplicación de SG en distintos estadios de un programa de mejoramiento genético de trigo para la predicción de caracteres múltiples, el manejo de la IGA y la selección de los mejores cruzamientos.

1.2.2. Objetivos específicos:

1. Mejorar la capacidad predictiva del valor genético de los individuos modelando la IGA en la predicción de nuevos individuos y de individuos parcialmente fenotipados.
2. Comparar la habilidad predictiva de modelos que incluyen caracteres múltiples correlacionados buscando optimizar la asignación de los recursos de fenotipado.
3. Predecir la varianza genética de la progenie considerando la población en equilibrio o desequilibrio de ligamiento y comparar la selección de los mejores cruzamientos incluyendo o no la varianza en dicha selección.

1.3. ESQUEMA GENERAL DE TESIS

En los primeros capítulos se presentan la introducción y el marco teórico de la tesis (Capítulo 1 y Capítulo 2). Se introducen los conceptos claves del mejoramiento genético de plantas autógamias y las estrategias de mejoramiento disponibles para obtener variedades mejoradas. Luego se presentan las herramientas moleculares actualmente disponibles para el mejoramiento, dentro de las cuales se incluye SG. Se introducen los conceptos de IGA, selección por caracteres múltiples y selección de cruzamientos para su aplicación en el contexto de SG.

Capítulo 3: En este capítulo se evalúa la capacidad de predecir el rendimiento con modelos que incluyen o ignoran la IGA en escenarios con distintos niveles de IGA. Estos modelos se evalúan en su capacidad predictiva del rendimiento en nuevas líneas experimentales y en líneas experimentales que no fueron evaluadas en los ambientes a

predecir. Se presenta el artículo 1 de la tesis: Lado B, Barrios PG, Quincke M, Silva P, Gutiérrez L. 2016. Modeling Genotype \times Environment Interaction for Genomic Selection with Unbalanced Data from a Wheat Breeding Program. *Crop Science*, 56:1–15.

Capítulo 4: En este capítulo se evalúa la capacidad predictiva de modelos multi-carácter de calidad panadera para la optimización de recursos de un programa de mejoramiento. Se evalúa cuanto del fenotipado de un carácter costoso y/o difícil de medir es posible reemplazar fenotipando uno o más caracteres correlacionados económicos y/o de simple medición. También se evalúan distintas estrategias de fenotipado de los caracteres correlacionados en la población. Se presenta el artículo 2 de la tesis: Lado B, Vázquez D, Quincke M, Silva P, Gutiérrez L. *Theoretical and Applied Genetics* (aceptado, 10/09/2018).

Capítulo 5: En este capítulo se comparan estrategias para la selección de cruzamientos evaluando si la inclusión de la varianza del cruzamiento es relevante en la selección de los mejores padres para rendimiento y calidad panadera. La selección incluyendo la varianza se evalúa con dos estrategias de predicción, considerando las líneas en equilibrio de ligamiento o asumiendo el desequilibrio de ligamiento. El objetivo de este capítulo es definir una estrategia de predicción de cruzamientos. Se presenta el artículo 3 de la tesis: Lado B, Battenfield S, Guzmán C, Quincke M, Singh RP, Dreisigacker S, Peña RJ, Fritz A, Silva P, Poland J, Gutiérrez L. 2017. Strategies for Selecting Crosses Using Genomic Prediction in Two Wheat Breeding Programs. *The Plant Genome*, 20(2):1–12

Capítulo 6 y 7: Discusión general de la tesis y consideraciones finales.

2. MARCO TEÓRICO

2.1. MEJORAMIENTO GENÉTICO VEGETAL

El mejoramiento genético de cultivos tiene como objetivo obtener cultivares mejor adaptados al ambiente y con mejores rendimientos que sean estables en el tiempo. Para lograr esto, el mejoramiento utiliza la selección natural y artificial buscando reunir en un único material varias características favorables a una región y tiempo determinados (Allard 1980). Para esto, las principales herramientas para generar diversidad genética son: la hibridación, la mutagénesis y la transgénesis. Luego, las estrategias usadas para seleccionar los mejores individuos y obtener individuos estables que puedan ser usados para la producción de semillas en muchas generaciones son distintas para especies autógamas, alógamas o de propagación vegetativa (Fehr 1987).

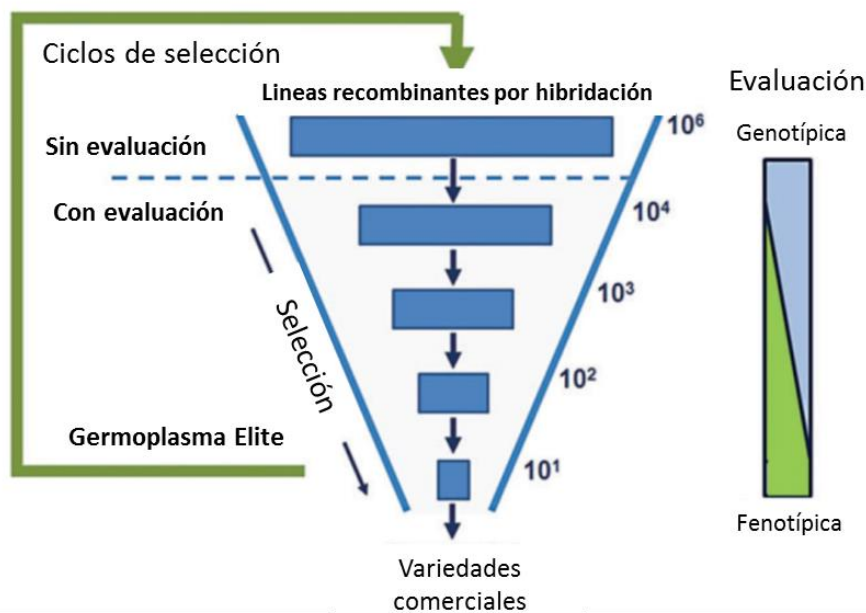


Figura 1: Esquema de un programa de mejoramiento donde se observa el proceso de endocria de líneas híbridas de maíz, en un esquema muy similar al de un programa de mejoramiento de plantas autógamas. Se muestra como comúnmente las nuevas líneas recombinantes provienen de líneas elite del programa. Al inicio hay muchas líneas para evaluar por lo que la evaluación fenotípica es mínima, a medida que el programa avanza se descartan líneas y las que quedan son líneas con mayor porcentaje de loci

homocigotos que son evaluadas fenotípicamente. Al final de un programa se incrementa la información fenotípica disponible para decidir entre las líneas experimentales que se convertirán en futuras variedades (adaptado de Cooper et al. (2014)).

2.1.1. Mejoramiento genético de plantas autógamias

La primera etapa de un programa de mejoramiento vegetal es la selección de los padres a cruzar para generar poblaciones segregantes por hibridación. En esta etapa se definen: los caracteres más importantes por los cuales se va a mejorar, la ganancia genética que será posible obtener y la diversidad genética, dada por el cruzamiento, con la que se contará durante el proceso de mejoramiento. Lo más común es seleccionar los padres entre los cultivares actualmente sembrados en la región y líneas elite del programa de mejoramiento, a menos que se necesite mayor diversidad (ej. resistencia a enfermedades o que no se cuente con materiales en la región). La ventaja de seleccionar como padres a los cultivares o las líneas elite del programa es que presentan características deseadas para la mayoría de los caracteres, son líneas adaptadas y con rendimiento estable (Bertan et al. 2007). Estas características de los padres son importantes para incrementar la probabilidad de obtener ganancias genéticas en la progenie. A su vez, debe tenerse en cuenta que los padres seleccionados provengan de ancestros diversos, para evitar disminuir la diversidad genética del programa (Fehr 1987).

En un programa de mejoramiento tradicional una forma de incrementar la variabilidad genética disponible es realizar cruzamientos de líneas elite del programa o cultivares actualmente sembrados en la región, con cultivares sembrados en otras regiones, padres más antiguos que son parte del material genético del programa, variedades adaptadas a la región, así como también realizar cruza inter-específicas (McCouch 2004; Fu 2015). Estos métodos se utilizan principalmente para nuevos caracteres o para resistencia a estreses bióticos o abióticos (Witcombe et al. 2008; Zaïm et al. 2017). Sin embargo, un cruzamiento con padres muy alejados genéticamente que no estén adaptados a las condiciones de la región implica un riesgo de perder parte de la adaptación presente en los cultivares conocidos (Fehr 1987).

Una vez realizados los cruzamientos, la siguiente etapa dentro de un programa de mejoramiento de plantas autógamias consiste en obtener líneas puras, que sean homocigotas para muchos de sus alelos por autofecundación. Las líneas puras pueden obtenerse utilizando distintas estrategias entre las cuales se destacan los métodos bulk, masal, pedigrí, descendencia de semilla única (DSU) y haploides duplicados (HD) (Baenziger y Depauw 2009). Los métodos se diferencian en que algunos realizan selección y otro no, en el tipo de selección que realizan (selección natural y/o artificial), en la posibilidad de mantener las familias identificadas al avanzar generaciones y en el tiempo que requieren para la obtención de líneas puras. Las distintas metodologías pueden combinarse ofreciendo una mayor cantidad de opciones para el mejoramiento genético (Fehr 1987) (ver Tabla 1).

El método de bulk implica la cosecha en masa de las plantas F_2 de las cuales se siembra una muestra al azar de semillas F_3 cosechadas (Figura 2). Este proceso se repite hasta obtener el nivel de homocigosis deseado a partir del cual se cosechan plantas individuales (generalmente plantas F_5). Las plantas F_6 , derivadas de las plantas individuales F_5 , se evalúan fenotípicamente para su selección. El método permite la acción de la selección natural favoreciendo la producción de semillas de los individuos más vigorosos mientras que los menos vigorosos aportan menos semilla a la siguiente generación (Harlan y Martini 1938). El método de bulk es sencillo y requiere pocos recursos humanos, sin embargo se corre el riesgo de perder materiales interesantes si el ambiente de selección es desfavorable a un genotipo de interés (Spitters 1979).

El método de selección masal es similar al método de bulk con el agregado de selección artificial de plantas o semillas por parte del mejorador. El método masal implica la evaluación de la población segregante de líneas F_2 y las plantas que no cumplen con determinados requisitos (ej. altura) se eliminan y el resto de las plantas se cosechan juntas (Figura 2). Del total de las semillas F_3 cosechadas se siembra una muestra, que puede ser seleccionada al azar o con algún criterio específico (ej. tamaño de semilla). En las siguientes generaciones (plantas F_3 , F_4 y F_5) se repite el procedimiento hasta obtener plantas altamente homocigotas. En el método masal actúa la selección natural y artificial, ya que las plantas menos vigorosas están menos representadas en el bulk de semillas cosechado y a su vez se descartan plantas que no

cumplen con los requisitos del mejorador. Este método requiere pocos recursos humanos y es eficaz en incrementar la frecuencia del genotipo de interés en una población. La gran desventaja que presenta el método es que no es útil para seleccionar caracteres de baja heredabilidad y también se corre el riesgo de perder materiales de interés si se siembran en ambientes desfavorables para un determinado genotipo (Fehr 1987).

En el método de pedigrí se seleccionan plantas F_2 individuales que se siembran en hilera (surcos) de progenie. Luego se seleccionan las familias (surcos) y dentro de cada familia se seleccionan las mejores plantas F_3 . Luego las semillas provenientes de una planta F_3 seleccionada se siembran en una hilera y mientras que en hileras contiguas se siembran las semillas pertenecientes a otras plantas F_3 de la misma familia. Se seleccionan las mejores familias, luego las mejores hileras dentro de cada familia y finalmente las mejores plantas F_4 dentro de cada hilera-familia (Figura 2). Este proceso de selección planta-hilera-familia se repite hasta que las líneas sean fenotípicamente homogéneas. Las plantas homogéneas de la misma hilera se cosechan juntas en bulk (Love 1927). Este es un método en el cuál predomina la selección artificial desde etapas muy tempranas. La ventaja de este método es que permite seleccionar cuando existe la mayor variabilidad genética y toma en cuenta la diversidad genética entre líneas. Sin embargo, como la selección se realiza en base al fenotipo de pocas plantas, resulta un método útil para seleccionar solo por caracteres de alta heredabilidad. El método de pedigrí requiere muchos recursos humanos ya que es necesario sembrar muchas líneas en distintos ambientes dónde se maximice la expresión de los caracteres de interés (Fehr, 1987).

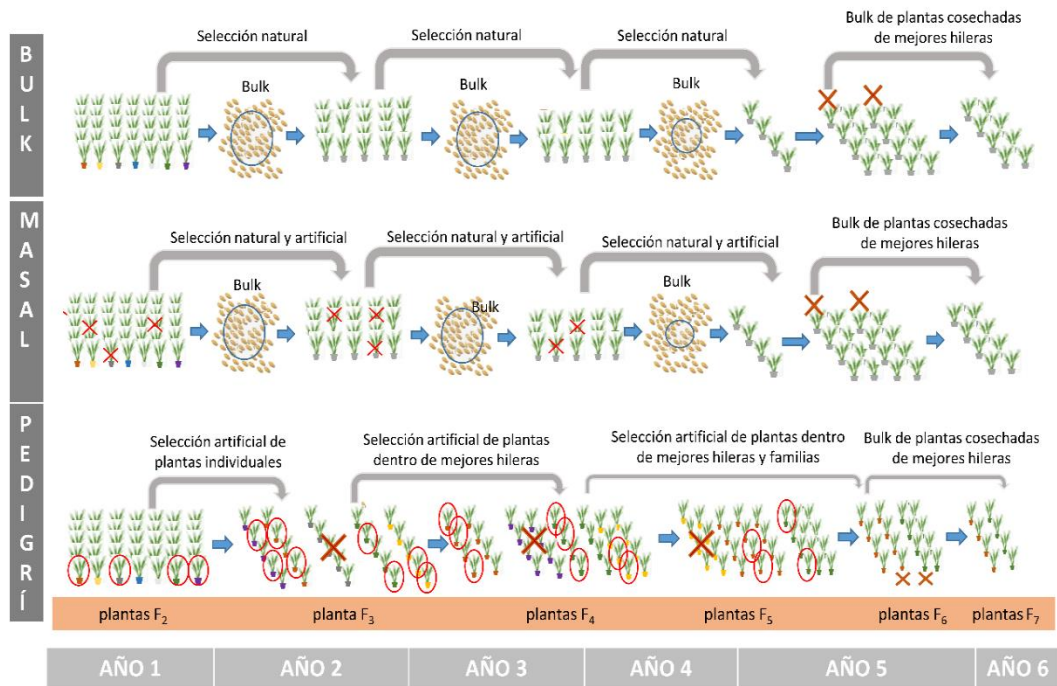


Figura 2. Métodos de avance generacional en los cuales hay selección durante el proceso de obtención de líneas puras.

El método de descendencia de semilla única es otra estrategia en la cual se avanza la progenie de las plantas F₂ a partir de semillas individuales, una por planta (Brim 1966). Las semillas F₃ cosechadas se siembran de forma individual y luego se cosecha una semilla F₄ por cada planta F₃ (Figura 3). Este proceso se repite hasta la obtención de líneas homocigotas. En este método no existe selección durante el proceso de avance de generaciones y todas las líneas F₂ estarán representadas en la generación final, excepto que exista algún problema con la viabilidad de las semillas (Fehr 1987). La gran ventaja de este método es que permite reducir el tiempo de desarrollo de líneas puras dado que, al cosechar una única semilla sin selección, pueden realizarse varios ciclos por año acortándose el ciclo de desarrollo de la planta en ambientes controlados. Se pueden realizar, de 2 a 6 generaciones por año en invernáculo o a contra estación (Snape y Riggs 1975; Watson et al. 2018). Las desventajas del método radican en que durante el proceso de avance de generaciones no se evalúan los desempeños de los genotipos y que fallas en la germinación o

supervivencia de una planta eliminan progenies completas (Baenziger y Depauw 2009).

Otro método de avance generacional para la obtención de líneas puras es el método de HD, que permite generar plantas haploides a partir de un cultivo de anteras o microesporas de plantas F_1 o por eliminación de cromosomas mediante cruzamientos inter-específicos (ej. con plantas de maíz). Los cromosomas de las plantas haploides regeneradas se duplican utilizando colchicina para obtener plantas diploides (Laurie y Bennett 1988; Guzy-Wrobelska y Szarejko 2003). Finalmente se cosechan las semillas de estas plantas diploides que son homocigotas para todos sus alelos (Henry y Buyser 1990). Durante la generación de HD no hay selección artificial ya que no es posible la evaluación fenotípica. Sin embargo, puede ocurrir un proceso de selección natural, ya que genotipos con producción preferencial de haploides pueden estar más representados que el resto, mientras que genotipos que no produzcan haploides se pierden. Es un método caro de aplicar a gran escala. Su mayor ventaja es que permite acelerar los tiempos de desarrollo de una línea homocigota en comparación con los métodos de bulk, masal y pedigrí (Fehr 1987).

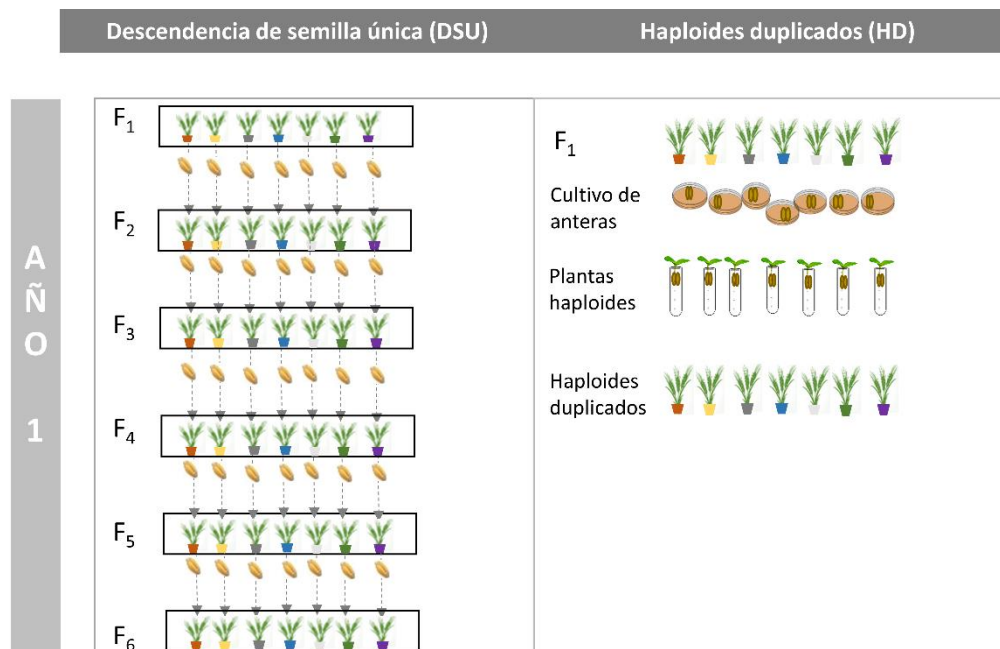


Figura 3. Métodos de avance generacional que no incluyen selección durante el proceso de obtención de líneas puras.

Independientemente de la combinación de métodos que se aplique, el éxito en la obtención de nuevos cultivares “mejorados” reside en identificar la mejor combinación genética de los padres o las cruzas que darán mejor descendencia y en identificar los individuos superiores en el menor tiempo posible. La identificación de las líneas experimentales superiores puede hacerse directamente por el fenotipo y/o indirectamente mediante el uso de marcadores.

Tabla 1. Características principales de los métodos de obtención de líneas puras

<i>Método</i>	<i>Bulk</i>	<i>Masal</i>	<i>Pedigrí</i>	<i>DSU</i>	<i>HD</i>
<i>Selección</i>	Natural	Natural y artificial	Artificial	Sin selección	Sin selección
<i>Evaluación fenotípica</i>	Tardía	Temprana	Temprana	Tardía	Tardía
<i>Riesgo de pérdida líneas</i>	Alto	Alto	Medio	Bajo	Bajo
<i>Requerimiento de recursos humanos</i>	Bajo	Bajo	Alto	Alto	Alto
<i>Costo por año</i>	Bajo	Bajo	Medio/Alto	Alto	Alto
<i>Tiempo (años)</i>	5-6	5-6	5-6	1-3	1-2
<i>Ambiente</i>	Favorable	Favorable	Favorable	Desfavorable (invernáculo)	Desfavorable (laboratorio)

La selección directa por el fenotipo de interés ha sido fundamental en la obtención de cultivares mejorados (Duvick 2005; Graybosch y Peterson 2010; Hawkesford et al. 2013). Entre varios aspectos, el éxito de dicha mejora se debe al desarrollo de diseños de experimentos que permiten diferenciar las distintas fuentes de variación involucradas en la expresión del fenotipo, discriminando la variación genotípica, de la variación ambiental y de la variación debida al error experimental (Fisher 1925; Federer 1955). Una forma de evaluar el éxito de la selección fenotípica es mediante el cálculo de la respuesta a la selección:

$$R = h^2 * i * \sigma_p$$

donde h^2 es la heredabilidad en sentido estricto; i es la intensidad de selección; y σ_p es la raíz cuadrada de la varianza fenotípica. La intensidad de selección y la varianza fenotípica estarán determinadas por el número de plantas seleccionadas y los cruzamientos realizados (Fehr 1987). La heredabilidad en sentido estricto se define

como la proporción de la varianza fenotípica atribuida a la varianza genética aditiva (σ_a^2/σ_p^2) (Visscher et al. 2008). La selección fenotípica de caracteres de muy baja heredabilidad es difícil de realizar porque la expresión del carácter depende en gran medida del ambiente. En estos casos, la selección utilizando la información molecular presenta mayor utilidad, ya que permite independizarla del ambiente y acortar tiempos en selección de líneas experimentales (Lande y Thompson 1990).

2.2. SELECCIÓN ASISTIDA POR MARCADORES

La selección asistida por marcadores (SAM) es una estrategia de selección indirecta de individuos utilizando marcadores moleculares que están asociados al fenotipo de interés (Lande y Thompson 1990; Bonnett et al. 2005).

2.2.1. Marcadores Moleculares

Existen algunas características que hacen que algunos marcadores sean preferidos frente a otros para su aplicación en la SAM. En general, se pretende que los marcadores tengan altos niveles de polimorfismos para que sean útiles en la diferenciación de distintas variedades, que sean co-dominantes para diferenciar plantas homocigotas de heterocigotas, que sean abundantes y con amplia distribución en el genoma, de forma de encontrar marcadores ligados a los genes de interés. Además, desde el punto de vista práctico, se pretende que sean fáciles de identificar en el laboratorio y de bajo costo para su amplia aplicación en muestras de un programa de mejoramiento (Jiang 2013).

Los primeros marcadores en ser desarrollados fueron los polimorfismos en el largo de los fragmentos de restricción (RFLP), inicialmente utilizados para la construcción del mapa de ligamiento humano (Botstein et al. 1980). Los RFLP son fragmentos de ADN producto de la digestión del ADN con enzimas de restricción. Los fragmentos se corren por electroforesis y se obtienen distintos perfiles de restricción. Los diferentes perfiles se deben a variantes en la digestión de los fragmentos de ADN que son atribuidos a cambios en el ADN (inserciones, deleciones, cambios de base, etc). Estos marcadores son co-dominantes, reproducibles y no requieren equipamiento especial. El problema que presentan es que los perfiles de RFLP pueden ser difíciles de interpretar sin conocimiento de la secuencia del fragmento que se está analizando,

tienen bajos niveles de polimorfismo por locus, son muy laboriosos para correr en muchas muestras, difíciles de automatizar y requieren mucho ADN de buena calidad (Gupta et al. 1999). Los RFLP han sido útiles en el mapeo de genes en plantas (Tanksley et al. 1989) aunque específicamente, su utilización ha sido limitada debido a su baja abundancia. Sin embargo, presentan la ventaja de distinguir entre locus en cromosomas homeólogos (Gupta et al. 1999).

Otro grupo de marcadores moleculares que han sido usados en el mejoramiento de plantas se basan en la amplificación por la técnica de reacción en cadena de la polimerasa (PCR). Dentro de este grupo los marcadores más destacados han sido: marcadores de amplificación al azar de ADN polimórfico (del inglés RAPD), marcadores de polimorfismo en el largo de los fragmentos amplificados (del inglés AFLP), los microsatélites o secuencias simples repetidas (SSR) y los polimorfismos de una sola base (SNP). Los RAPD son amplificaciones al azar del genoma realizada con oligonucleótidos diseñados de forma genérica, luego los fragmentos de ADN amplificados se evalúan en un gel de agarosa o acrilamida y se comparan los perfiles de amplificación entre distintos genotipos. Las variantes en los perfiles de amplificaciones puede deberse a cambios en el ADN, sustituciones, inserciones y/o deleciones. Estos marcadores tienen la ventaja sobre los RFLP en que su instrumentación es más sencilla y económica, sin embargo han sido menos usados, dada su baja reproducibilidad (Jiang 2013). Los AFLP son marcadores generados por una modificación de los RFLP en la cual se introduce la amplificación selectiva por PCR de los fragmentos obtenidos de la digestión con enzimas de restricción (Vuylsteke et al. 2007). Los AFLP presentan ventajas frente a los RFLP ya que son marcadores más fáciles de interpretar, requieren menos ADN, son menos laboriosos, menos costosos, pueden ser automatizados y son más abundantes en el genoma (Vuylsteke et al. 2007; Gupta et al. 2008; Jiang 2013). Sin embargo, presentan algunas limitaciones dado que requieren una alta calidad de ADN, son dominantes y no son útiles para detectar pequeñas variaciones, como ser cambio de una sola base (Vuylsteke et al. 2007). Otro tipo de marcadores son los SSR, que son repeticiones de secuencias de 1 a 6 pares de bases en el ADN, que varían en la cantidad de repeticiones de dicha secuencia. La técnica más usada para identificarlos requiere el desarrollo de

oligonucleótidos específicos que amplifiquen estas regiones (Gupta et al. 1999; Gupta y Varshney 2000). Los SSR son marcadores muy utilizados en mejoramiento de plantas, debido a que son altamente polimórficos, muy abundantes y presentan una buena distribución en el genoma (Condit y Hubbell 1991; Gupta y Varshney 2000). La desventaja de este tipo de marcadores es que requieren conocimiento de la secuencia a amplificar y que son laboriosos de desarrollar (Jiang 2013).

Los RFLP, AFLP y SSR han sido útiles en el mejoramiento genético vegetal para la construcción de mapas de ligamiento, mapeo, introducción de genes por retrocruzamientos y para selección de líneas promisorias en etapas tempranas (Varshney y Tuberosa 2007; Korzun et al. 1997; Kojima et al. 1998; Korzun et al. 1998; Ganal y Röder 2007; Silva et al. 2015). Sin embargo, son laboriosos y relativamente costosos para ser aplicados a gran escala y rutinariamente en un programa de mejoramiento (Thomson 2014).

Los polimorfismos de una sola base (SNP) son un tipo de marcador empleado más recientemente, que ha cobrado importancia debido a su abundancia en el genoma, a las tecnologías desarrolladas para su identificación, a la posibilidad de automatización que presentan y los bajos costos para su aplicación a gran escala (Gupta et al. 2008; Davey et al. 2011). Actualmente existen plataformas a diferentes escalas para identificar SNP. Las plataformas de pequeña escala están basadas en geles de agarosa o acrilamida y las de mediana escala están basadas en tecnologías de PCR en tiempo real (ej. TaqMan o KASP™). Las plataformas de gran escala se clasifican en plataformas fijas, basadas en métodos de hibridación (ej. microarreglos) o en plataformas *de novo* que utilizan métodos de re-secuenciación del genoma (ej. genotipado por secuenciación). El desarrollo de plataformas a gran escala es lo que le ha dado mayor difusión a este tipo de marcadores (Thomson 2014).

Tabla 2. Propiedades de los marcadores moleculares más utilizados en selección asistida por marcadores.

Marcadores	RFLP	RAPD	AFLP	SSR	SNP
Técnica de detección	Enzimas de restricción y geles	PCR y geles	PCR y geles	PCR y geles	PCR y geles, PCR-RT, microarreglos y re-secuenciación
Dominancia	Co-dominante	Dominante	Dominante	Co-dominante	Co-dominante
Polimorfismo en el genoma	Bajo	Medio	Medio	Alto	Alto
Reproducibilidad	Alta	Baja	Alta	Alta	Alta
Abundancia	Baja	Media	Media/Alto	Alta	Muy Alta
Grado de labor	Alta	Baja	Media	Media - baja	Muy Baja
Grado de automatización	Bajo	Moderado	Moderado	Alto	Muy Alto

Dentro de las plataformas a gran escala, las fijas fueron las primeras en desarrollarse y se conocen como microarreglos de ADN o chips (ej. BeadArray y Affymetrix). Se inmovilizan fragmentos de ADN sobre plataformas sólidas y luego el ADN marcado, proveniente de las muestras, se une al ADN fijado sobre la plataforma, si la secuencia es exactamente complementaria (Wang et al. 1998; Rasheed et al. 2017). Las ventajas de estas plataformas es que son flexibles para genotipar un amplio rango de muestras y número de marcadores. A su vez, son sencillas y rápidas de aplicar a gran escala. La mayor desventaja es el costo de desarrollo de esta tecnología y el hecho de que los microarreglos desarrollados para un programa de mejoramiento no son 100% transferibles a otro programa dado que muchos SNP serán monomórficos (Thomson 2014).

Estas limitantes han conducido al desarrollo de tecnologías de secuenciación *de novo*, que permiten el genotipado de miles de marcadores en cientos a miles de individuos, sin la necesidad de una identificación previa de los marcadores (Hyten et al. 2010). Dentro de estas nuevas tecnologías se encuentran los métodos que comprenden la preparación de librerías, secuenciación y posterior análisis bioinformático para identificar los SNPs, los cuales se basan en la re-secuenciación del genoma en pequeños fragmentos de ADN de las distintas muestras a genotipar

(Davey et al. 2011). La alineación de estos fragmentos entre sí o contra un genoma de referencia, permite identificar variaciones genéticas de cambios en una única base nucleotídica, inserciones y deleciones de las mismas (Kumar et al. 2012a). 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; Lado et al. 2013). La mayor limitante que presentan estas tecnologías es la inversión inicial, por lo que en general, estas plataformas no son usadas directamente por programa de mejoramiento sino que se contratan como un servicio. Como consecuencia, se dificulta la logística de obtención de los marcadores al momento de realizar la selección. A su vez, para que estas tecnologías sean accesibles se secuencian un pool de varias muestras (ej. 96, 192, 384-plex) y se descartan regiones del genoma que no son de interés, esto genera mucha información genotípica (miles de SNP) incompleta, o sea con una alta proporción de datos faltantes. A su vez estas técnicas presentan pasos de amplificación por PCR previos a la secuenciación, que junto con los errores mismos de secuenciación inducen a errores de genotipado (Davey et al. 2011; Keschull y Zador 2015). Sin embargo, algunas de estas limitaciones se deben a que aún es reciente el desarrollo de estas tecnologías (Thomson 2014).

Existen diversas técnicas de preparación de librerías que utilizan enzimas de restricción, siendo las principales, 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). 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 de fragmentos, para la posterior secuenciación de los fragmentos seleccionados (Figura 4).

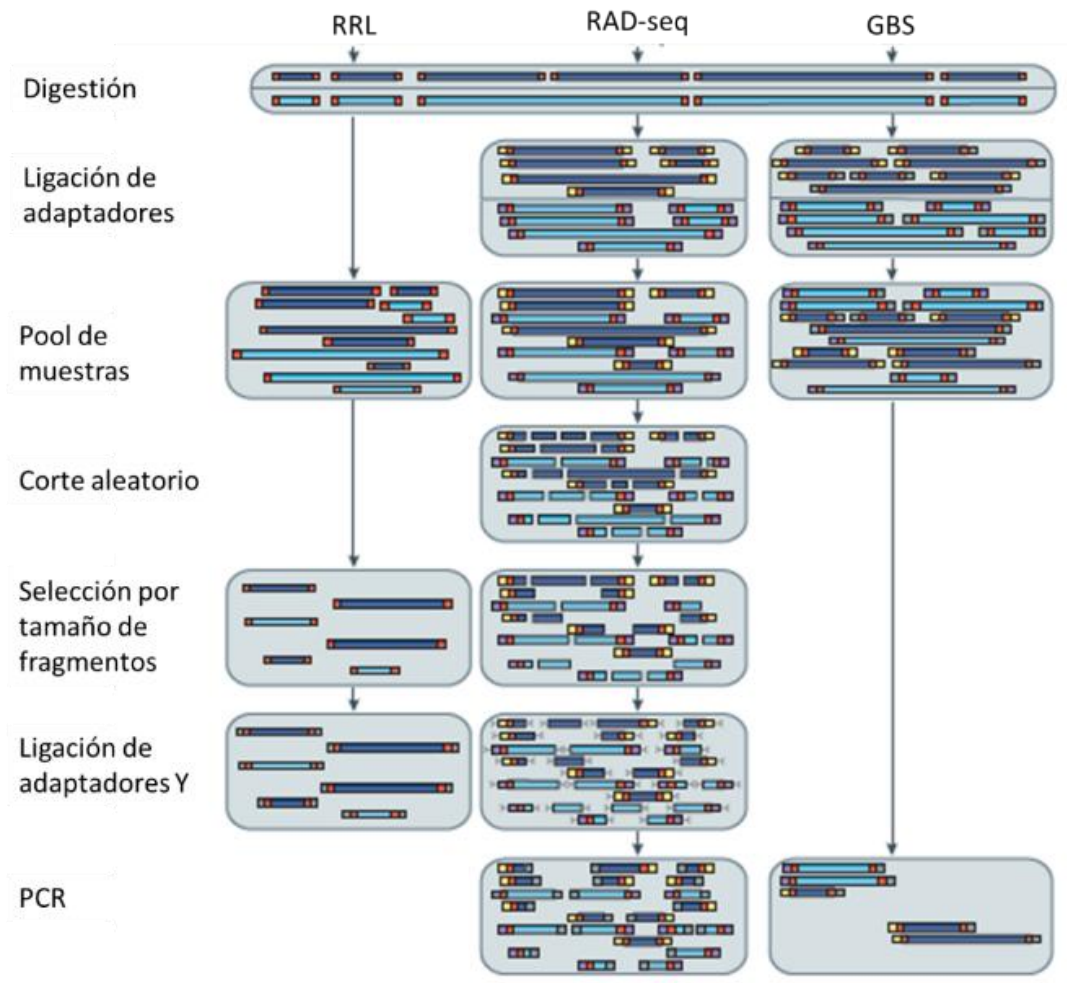


Figura 4. Esquema de la preparación de librerías de ADN por distintos métodos que utilizan enzimas de restricción, librerías de representación reducida (RRL), secuenciación de ADN asociada a sitios de restricción (RAD-seq) y genotipado por secuenciación (GBS) (adaptado de Davey et al. (2011)).

Una de las técnicas más populares de genotipado a gran escala es el genotipado por secuenciación (GBS). El GBS es un método de secuenciación de baja cobertura ya que los SNP obtenidos no están presentes en todas las muestras, por lo que los SNP identificados tienen muchos datos faltantes (Davey et al. 2011). Una ventaja muy importante de este método es su sencillez y el bajo costo por muestra (Elshire et al. 2011). El ADN digerido con enzimas de restricción se liga a adaptadores u oligonucleótidos que tienen un código de barras (secuencia de bases única) diferencial para cada muestra. Luego es posible hacer un pool de 96, 192 o 384 muestras que se

amplifica por PCR y se secuencian conjuntamente (Figura 5). Esta tecnología de genotipado se ha aplicado con éxito en muchas especies de plantas, por ejemplo en maíz (Elshire et al. 2011), switchgrass (Grabowski et al. 2014), zanahoria (Mezghani et al. 2018), papa (Uitdewilligen et al. 2013), cebada (Elshire et al. 2011), arroz (Huang et al. 2009), garbanzos (Kujur et al. 2015), colza (Bayer et al. 2015), sorgo (Morris et al. 2013) y trigo (Poland y Rife 2012; Lado et al. 2013), entre otras.

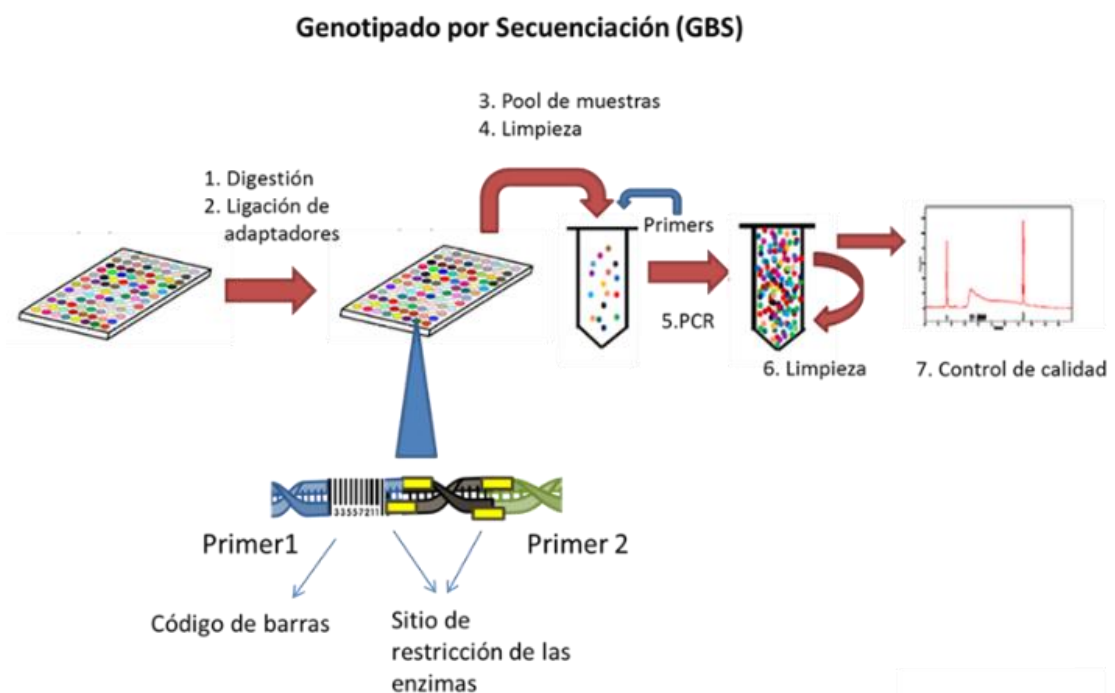


Figura 5. Esquema de la metodología de genotipado por secuenciación (GBS), en el cual se indica detalladamente los sucesivos pasos de la técnica (adaptado de Elshire et al. (2011)).

2.2.2. Mapeo de QTL

La identificación de marcadores asociadas a regiones genómicas involucradas en la expresión del carácter cuantitativo de interés se realiza mediante mapeo de loci cuantitativo (análisis de QTL; Tanksley 1993 y mapeo asociativo (GWAS); Jannink y Jansen 2001). El análisis de QTL cobró importancia entre la década de 1980 y 1990 con los primeros estudios en tomate y maíz (Paterson et al. 1990; Stuber et al. 1992). Esta metodología fue ampliamente aplicada a distintos cultivos, debido a la identificación de regiones genómicas involucradas en la expresión de caracteres

complejos (Andaya y Mackill 2003; Kirigwi et al. 2007; Cuesta-Marcos et al. 2009; Hayes et al. 2013). Este tipo de análisis requiere el desarrollo de poblaciones segregantes y balanceadas que deben ser genotipadas y fenotipadas, y contar o construir un mapa de ligamiento para la población a analizar (Tanksley 1993; Lee 1995). A su vez, para su aplicación en el marco de un programa de mejoramiento, los QTL identificados 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 mejores individuos implica identificar los alelos de los marcadores ligados a los QTL, en cientos a miles de individuos. Una de las limitantes del análisis de QTL es el corto historial de recombinación en las poblaciones desarrolladas, por lo que es un análisis que tiene bajo poder de detección cuando los marcadores asociados a QTL están en zonas del cromosoma con baja tasa de recombinación (Dole y Weber 2007). A su vez, este análisis sub-estima el número de QTL identificados y sobre-estima la magnitud del efecto de los QTL de efectos menores, esto se conoce como efecto Beavis (Xu 2003). Por último, otra limitante de este análisis es la dificultad para estimar las interacciones entre marcadores (epistasia), dado que no es posible incluir las interacciones en el modelo porque incrementa en gran medida el número de parámetros a estimar (problema $p \gg n$).

Como alternativa al análisis de QTL en poblaciones específicas se desarrolló el mapeo asociativo del genoma completo (GWAS), que permite identificar QTL en poblaciones diversas (familias, líneas endocriadas, etc.), como son las disponibles en un programa de mejoramiento. Una de las dificultades presentes en GWAS es el manejo de la estructura de la población, asegurando que las asociaciones entre los marcadores y los fenotipos se deben a un ligamiento físico entre los marcadores y los QTL involucrados en la expresión del fenotipo. Se han desarrollado diversas estrategias para considerar esta estructura en los modelos y eliminar los falsos positivos (Pritchard et al. 2000; Yu et al. 2006; Malosetti et al. 2007). El GWAS ha sido aplicado con éxito en la identificación de QTL en poblaciones de mejoramiento (Gutiérrez et al. 2011; von Zitzewitz et al. 2011; Locatelli et al. 2012). A pesar del éxito de GWAS, el efecto Beavis sigue siendo una limitante al identificar QTL de variables complejas. A su vez, los QTL identificados para caracteres complejos solo

explican una proporción de la varianza fenotípica, fenómeno que es conocido como heredabilidad perdida. Las causas de dicho fenómeno son múltiples y entre las principales se encuentran factores metodológicos (tamaño poblacional, número de marcadores, etc.), variantes alélicas raras en la población e interacciones génicas difíciles de detectar (Manolio et al. 2009).

La identificación de genes y el desarrollo de marcadores han dado lugar a la aplicación de SAM. Actualmente existen plataformas con marcadores disponibles para aplicar SAM en distintos caracteres de interés (ej. <http://maswheat.ucdavis.edu/>). La SAM ha sido exitosa en la selección de caracteres influenciados por pocos genes e introgresión de genes en variedades de interés agronómico (Hospital et al. 1997; Eathington et al. 2007; Ribaut y Ragot 2007; Silva et al. 2015), como es el caso de los genes de semi-enanismo en trigo (Ellis et al. 2002). Sin embargo, presenta dificultades al aplicarse en caracteres complejos influenciados por varios genes (Dekkers 2004). Las características de mayor importancia agronómica, como ser rendimiento, calidad y resistencia a enfermedades, presentan en general herencia compleja o cuantitativa (Mackay et al. 2009). Dichos caracteres complejos están influenciado por muchos genes que tienen acción pleiotrópica sobre el carácter estudiado (Kearsey y Farquhar 1998). A estas complejidades se le suman otros tipos de acciones génicas como ser dominancia completa, dominancia incompleta y epístasis (Carey 2000).

2.3. SELECCIÓN GENÓMICA

La SG es una herramienta de selección indirecta por marcadores moleculares de los individuos en base a su valor genético. El valor genético de un individuo es la suma de los efectos de los alelos presentes en el mismo (Meuwissen et al. 2001). Los efectos de los alelos se estiman utilizando modelos estadísticos que reúnen la información genotípica y fenotípica de un grupo de individuos que se conocen como población de entrenamiento. Estos efectos alélicos estimados son utilizados para predecir los valores genéticos de individuos que solo cuentan con información genotípica (Meuwissen et al. 2001; Heffner et al. 2011). Otra forma de predecir el valor genético de un individuo es utilizando modelos que incluyen la información fenotípica de la población de entrenamiento junto con la información de relacionamiento genético de dicha

población con el individuo desconocido a predecir (Hayes et al. 2009). La información de relacionamiento genético se calcula a partir de los marcadores moleculares, obteniéndose matrices de similitud genética que son usadas como matrices de varianza-covarianza entre los individuos (VanRaden 2008). A diferencia del análisis de QTL y GWAS, en los cuales solo se consideran los marcadores con efectos mayores, en la SG se consideran los efectos de todos los marcadores para predecir el valor genético de los individuos (Heffner et al. 2009). Utilizando SG se espera capturar regiones genómicas con menor efecto en el fenotipo, las cuales se enmascaran en las estrategias de análisis de QTL (Goddard y Hayes 2009). La SG surgió y se desarrolló inicialmente en el mejoramiento animal, debido a que previo al desarrollo de tecnología de genotipado, en animales ya se realizaba la selección de individuos en base a sus valores de cría obtenidos utilizando matrices de pedigrí como matrices de varianza-covarianza entre individuos (Hayes et al. 2009; Nakaya y Isobe 2012). Luego, junto con el desarrollo de técnicas de genotipado a gran escala, la SG comenzó a aplicarse con mucho éxito en el mejoramiento vegetal (Heffner et al. 2009; de los Campos et al. 2009b; Crossa et al. 2010b; Iwata y Jannink 2011; Lorenz et al. 2011; Heslot et al. 2015; Bassi et al. 2016; Michel et al. 2016).

2.3.1. Modelos de predicción

Los modelos estadísticos utilizados para predecir los valores genéticos se clasifican según la estrategia estadística usada en: modelos de regularización, modelos de selección de variables y modelos de kernels, entre otros (Meuwissen et al. 2001; Gianola et al. 2006; Solberg et al. 2008; Gianola y van Kaam 2008; de los Campos et al. 2009a; Endelman 2011; Gianola et al. 2011). El problema que intentan solucionar estos modelos es que el número de predictores (efecto de los marcadores) exceden el número de observaciones ($p \gg n$). Por lo tanto no existe una solución única en la estimación del efecto de todos los marcadores en simultáneo por el método clásico de mínimos cuadrados, debido a la existencia de más incógnitas que ecuaciones normales (Endelman y Jannink 2012).

Los modelos lineales que utilizan técnica de penalización o regularización solucionan el problema de $p \gg n$, penalizando las estimaciones de los coeficientes de regresión (efecto de los marcadores) mediante la inclusión de un parámetro lambda

que produce una contracción de los coeficientes entorno a cero (Hoerl y Kennard 1970). Así, se reduce la co-linealidad (redundancia) entre los predictores (marcadores). El parámetro lambda puede estimarse minimizando el error del modelo siguiendo la misma estrategia que en mínimos cuadrados:

$$[1] \quad y = X\beta_{RR} + \varepsilon_{RR} \quad (\text{Ridge Regression})$$

donde y es el vector fenotípico; X es una matriz de incidencia de los efectos fijos que contiene la información que relaciona los marcadores con los valores fenotípicos; β es el vector de efectos fijos (media general y efecto de los marcadores); ε es el vector de efectos aleatorios de los residuos que siguen una distribución normal con $E(\varepsilon_{RR}) = 0$; $var(\varepsilon_{RR}) = R$, donde generalmente:

$$R = I\sigma_{\varepsilon_{RR}}^2.$$

Los efectos de los marcadores se estiman de la siguiente forma:

$$\hat{\beta}_{RR} = (X'X + \lambda I)^{-1} X' y$$

En el modelo [1] el parámetro β estimado se diferencia del método de mínimos cuadrados por la inclusión del parámetro λ , que cuanto mayor sea, mayor será la contracción sobre los efectos de los marcadores (Whittaker et al. 1999).

Otra formas de resolver estos modelos es aplicando penalización en el contexto de los modelos mixtos. Cuando se estima el efecto de los marcadores, se asume que los efectos provienen de una distribución normal, con media cero y varianza común a todos los marcadores (*Ridge Regression Best Linear Unbiased Prediction*; RR-BLUP). Cuando se predicen directamente los valores de cría de los individuos, se asume que los valores de los individuos provienen de una distribución normal con media cero y una matriz de varianza-covarianza entre individuos estimada con los marcadores moleculares (*Genomic Best Linear Unbiased Prediction*, G-BLUP) (Henderson 1975; Piepho 2009).

El modelo RR-BLUP se representa como:

$$[2] \quad y = X\beta_{blup} + ZMu_{blup} + \varepsilon_{blup} \quad (\text{RR-BLUP})$$

donde y es el vector fenotípico, X es la matriz de incidencia de efectos fijos y β_{blup} el vector de efectos fijos; Z es una matriz de incidencia que une los individuos con las observaciones; M es la matriz de individuos por marcadores codificada como (-1,0,1);

u_{blup} , ε_{blup} son vectores aleatorios con el efecto de los marcadores y de los residuos, respectivamente. Los residuales se distribuyen igual que en el modelo [1] y u se distribuye normal con $E(u) = 0$; $var(u_{blup}) = G$, siendo $G = I\sigma_u^2$.

Los efectos fijos y aleatorios se estiman utilizando las siguientes ecuaciones:

$$\hat{u}_{blup} = GM'Z'V^{-1} (y - X\hat{\beta}_{blup})$$

$$\hat{\beta}_{blup} = (X'V^{-1}X)^{-1}X'V^{-1}Y$$

$$V = ZMG M'Z' + R$$

Estimar estos parámetros requiere invertir la matriz V , esto puede ser computacionalmente tedioso por lo que se usan las ecuaciones de Henderson (Henderson 1975):

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}ZM \\ M'ZR^{-1}X & M'Z'R^{-1}ZM + G^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ u \end{pmatrix} = \begin{pmatrix} X'R^{-1}y \\ M'Z'R^{-1}y \end{pmatrix}$$

Si $G = \sigma_u^2 I$ y $R = \sigma_e^2 I$

$$\begin{pmatrix} \beta \\ u \end{pmatrix} = \begin{pmatrix} X'X & X'ZM \\ M'Z'X & M'Z'ZM + \lambda I \end{pmatrix}^{-1} * \begin{pmatrix} X'y \\ M'Z'y \end{pmatrix}$$

Resolviendo esta ecuación estimamos el valor de los parámetros, λ es el inverso de la heredabilidad ($\lambda = \sigma_e^2 \sigma_u^{-2}$) y es el parámetros de penalización. En la ecuación de estimación de \hat{u} es posible ver como la diferencia entre el valor estimado y la media ($y - X\hat{\beta}$) se penaliza en función de la covarianza genética entre las líneas y la varianza total estimada ($GM'Z'V^{-1}$). El método más común para estimar los parámetros del modelo [1] es máxima verosimilitud restringida (REML).

Los modelos mixtos más utilizados en SG son RR-BLUP y G-BLUP. El modelo RR-BLUP, como se mencionó anteriormente, estima los valores de cría (g) mediante la estimación del efecto de los marcadores, como $g = Mu_{blup}$. En cambio, el modelo G-BLUP utiliza el modelo típicamente usado en selección animal tradicional pero incluyendo la matriz de distancia genética estimada con los marcadores como matriz de varianza-covarianza de los individuos a predecir (Henderson 1984; VanRaden 2008).

El modelo G-BLUP puede escribirse como:

$$[3] \quad y = X\beta_{blup} + Zg_{blup} + \varepsilon'_{blup} \quad (\text{G-BLUP})$$

donde $y, X, \beta_{blup}, Z, \varepsilon'_{blup}$ se definen igual que en el modelo [2]; g_{blup} es el vector aleatorio de efecto de las líneas, g se distribuye normal con $E(g_{blup}) = 0$ y $var(g_{blup}) = K\sigma_g^2$; K es la matriz de relacionamiento genético estimada como $K = MM' \sigma_u^2 / \sigma_g^2$, siendo M la matriz de marcadores previamente definida.

Ambos modelos, GBLUP y RR-BLUP, son equivalentes como se presenta a continuación (Hayes et al. 2009; Endelman 2011):

$$E(g_{blup}) = 0$$

$$E(Mu_{blup}) = E(u_{blup}) = 0$$

$$Var(g_{blup}) = K\sigma_g^2 = MM' \sigma_u^2$$

$$Var(Mu_{blup}) = M * Var(u_{blup}) * M' = M I \sigma_u^2 M' = MM' \sigma_u^2$$

En los modelos anteriores (modelo [1] y [2], y por equivalencia el modelo [3]), los efectos de los marcadores son igualmente *encogidos* en torno a cero, asumiendo un modelo infinitesimal en el cual casi todos los marcadores aportan su “efecto” en el cálculo de los valores de cría (de los Campos et al. 2009b).

Si generalizamos la solución del modelo de *Ridge Regression* (modelo [1]) de la forma:

$$\hat{\beta} = \underset{\text{argmin}}{\left\{ \sum_i (y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p \|\beta_j\|^\gamma \right\}}$$

La primera parte del argumento indica la solución que minimiza los errores del modelo (solución de mínimos cuadrados ordinarios) y el segundo término es el de penalización, cuando λ es cero la solución es la de mínimos cuadrados (Frank y Friedman 1993).

Distintos valores de gamma (γ) confieren distintas formas de penalizar el modelo; cuando $\gamma = 2$, el modelo de penalización es el modelo de *Ridge Regression* (RR; modelo 1) (Hoerl y Kennard 1970). Cuando γ tiende a cero el modelo de penalización se transforma en un modelo de selección de variables, dado que el modelo

fuerza a que haya muchos marcadores con efecto cero. Cuando $\gamma = 1$, el modelo se conoce con el nombre de ‘Lasso’, correspondiendo a un modelo que mezcla la penalización y la selección de variables, dado que fuerza a que un mayor número de marcadores sean cero (en comparación con RR), mientras que los efectos de los marcadores que son distintos de cero son penalizados (Tibshirani 1996; de los Campos et al. 2013).

A su vez, algunas de las estrategias bayesianas usadas en la SG permiten combinar técnica de regularización con estrategias de selección de variables a través de la elección de la distribución *a priori*. Estas distribuciones en general, permiten que muchos marcadores tengan un efecto cero o casi cero y que pocos marcadores tengan un gran efecto. Los primeros métodos desarrollados fueron Bayes A y Bayes B (Meuwissen et al. 2001; de los Campos et al. 2009b) los cuales agregan la posibilidad, de considerar distintos *encogimientos* para los distintos marcadores, que no dispone RR-BLUP. Bayes A asume que el efecto de cada marcador es aleatorio, y el mismo proviene de una distribución *a priori* que es normal con media cero y su propia varianza. Por ende cada marcador se *encoge* diferencialmente entorno a cero. Bayes B utiliza una distribución *a priori* diferente de la normal, con mayor densidad de probabilidad en cero. Esto permite que un mayor número de marcadores tengan un efecto igual a cero y por lo tanto, que se seleccionen variables. Existen otros modelos bayesianos (ej. Bayes C π y Bayes D π) que buscan aproximar los supuestos del modelo a la realidad. De esta forma, intentan entre otras cosas, que las distribuciones *a priori*, sean más o menos informativas, logrando en mayor medida, que se utilice la información presente en los datos en la estimación de los parámetros (Yi y Xu 2008; Gianola et al. 2009).

Por otro lado, los métodos semi-paramétricos (Gianola et al. 2006) incorporan a los modelos anteriormente descritos, una función no-paramétrica con el objetivo de modelar las interacciones entre los marcadores. Esto libera al modelo del supuesto de linealidad entre los marcadores y el fenotipo. Al incluir las funciones no-paramétricas en un modelo mixto, se define un modelo llamado *Reproducing Kernel Hilbert Space* (RKHS), y pueden representarse como:

$$[4] \quad y = X\beta + Zg + k(M) + \varepsilon$$

donde en el vector $y, X, \beta, Z, g, \varepsilon$ son iguales al modelo G-BLUP y $k(M)$ es la función que relaciona los marcadores con el fenotipo. Existen diferentes funciones $k(M)$ que dan al modelo flexibilidad para tomar en cuenta relaciones aditivas, de dominancia y epítasis.

Las relaciones no aditivas entre marcadores pueden modelarse de forma similar a la usada en los modelos RKHS, utilizando el modelo G-BLUP, pero cambiando la forma de estimar la matriz de relaciones genéticas (G), considerando las relaciones no aditivas entre marcadores (Gianola et al. 2006). Endelman (2011) utiliza un *kernel* o núcleo Gaussiano de la forma:

$$G_{ij} = \exp[-(D_{ij}/\theta)^2]$$

donde D es la matriz de distancias genéticas euclidianas calculada a partir de los marcadores; θ es un parámetro de escala que regula cuán rápido se reducen las covarianzas genéticas con la distancia; y G es una matriz de varianza-covarianza que permite modelar las relaciones epistáticas entre marcadores a través del cálculo de las relaciones genéticas entre individuos.

Existen muchos trabajos que comparan modelos para realizar las predicciones pero hasta al día de hoy, no se ha identificado un modelo que sea mejor a otro en todos los casos (Meuwissen et al. 2001; Habier et al. 2007; Zhong et al. 2009; Lorenz et al. 2011; Heslot et al. 2012). El modelo más utilizado en plantas es el modelo G-BLUP dado su simplicidad y robustez (Heslot et al. 2015), aunque varios autores señalan la importancia de seguir explorando este campo en la búsqueda de modelos cuyas asunciones se ajusten más a la realidad (Heslot et al. 2015; Crossa et al. 2017), como por ejemplo considerar un encogimiento diferencial sobre el efecto de los marcadores según exista o no desequilibrio de ligamiento con algún QTL e incorporar interacciones génicas (Zhong et al. 2009; Lorenzana y Bernardo 2009; Lorenz et al. 2011).

2.3.2. Población de entrenamiento

La población de entrenamiento se utiliza para entrenar modelos de predicción o lo que es lo mismo, estimar los parámetros del modelo necesarios para predecir el comportamiento de nuevos individuos. El éxito de las predicciones va a depender de

la calidad y cantidad de los datos fenotípicos y genotípicos de esta población, así como del parentesco entre la población de entrenamiento y de predicción.

En relación a la calidad de los datos fenotípicos, es importante contar con una evaluación fenotípica precisa de esta población en distintos ambientes. Actualmente, la precisión en la evaluación fenotípica es una de las limitantes en la mejora de la habilidad predictiva de los modelos usados en SG (White et al. 2012; Cobb et al. 2013). Para superar esta limitante, se trabaja incluyendo nuevos datos fenotípicos que puedan ser medidos a gran escala con mayor precisión (Cooper et al. 2014; Rutkoski et al. 2016; Crain et al. 2018). A su vez, existe una nueva concepción en la SG basada en que los datos serán utilizados para estimar efectos de alelos. Por lo tanto, lo más importante sería realizar diseños de experimentos que estimen adecuadamente el efecto de los alelos. Este cambio podría modificar los diseños experimentales dónde se consideraran repeticiones alélicas en vez de repeticiones de individuos (Endelman et al. 2014).

En relación a los datos genotípicos, se debe conocer el tamaño efectivo de la población, el cual se relaciona con la tasa de endocria de la población en estudio. Las especies endocriadas como el trigo, presentan mayores distancias en la caída del desequilibrio de ligamiento que especies de polinización cruzada, es decir que presentan amplias regiones del genoma que se heredan juntas de generación en generación. En una población con mayor distancias en la caída del desequilibrio de ligamiento, es posible capturar los efectos de todos los QTL involucrados en la expresión del fenotipo utilizando menos densidad de marcadores que en poblaciones con menores distancias en la caída del desequilibrio de ligamiento. De todas formas, el número de marcadores actualmente no es una limitante en la mejora de la predicción de los valores genéticos de los individuos. Existen trabajos que demuestran un *plateau* en el incremento de la precisión de las predicciones al incrementar el número de marcadores (Asoro et al. 2011; Lorenz et al. 2011).

Por otro lado, el tamaño y la representatividad de la población de entrenamiento es determinante para la predicción de los valores genéticos de los individuos. La precisión de las predicciones mejora bajo dos condiciones; cuando la población de entrenamiento representa la variación fenotípica realmente existente para el carácter

de interés, y cuando ésta se relaciona genéticamente con la población de predicción, ya que entonces es posible estimar el efecto de todos los alelos presentes (Asoro et al. 2013; Hickey et al. 2014; Crossa et al. 2014; Isidro et al. 2015). Al igual que con el incremento en el número de marcadores, se ha reportado que al incrementar el número de individuos de la población de entrenamiento las precisiones del cálculo de los valores genéticos de los individuos aumenta hasta un valor determinado, a partir del cual el incremento de la población no produce estimaciones más precisas (Asoro et al. 2011; Akdemir et al. 2015; Isidro et al. 2015).

Para aplicar la SG de rutina en un programa de mejoramiento, es necesario contar con una estrategia de definición de la población de entrenamiento. Actualmente, existen algoritmos de optimización, que permiten seleccionar una determinada población de entrenamiento para predecir una población de predicción dada (Rincent et al. 2012; Akdemir et al. 2015; Isidro et al. 2015; Rincent et al. 2017). En estos algoritmos, la mejor población de entrenamiento se obtiene minimizando la varianza del error de predicción o maximizando el coeficiente de determinación entre la población de entrenamiento y la de predicción.

2.4. INTERACCIÓN GENOTIPO POR AMBIENTE

Una vez definidos los individuos que forman parte de la población de entrenamiento, es necesario que estas líneas sean caracterizadas con buena calidad de datos fenotípicos para las variables que se quieren predecir (Prasanna et al. 2013). Una buena caracterización para variables fenotípicas de baja heredabilidad, implica la determinación del carácter de interés en distintos años y ambientes. El análisis de datos provenientes de ensayos multi-ambiente (MET) incluye el manejo de la IGA.

La IGA se define como la diferencia en el desempeño relativo de dos o más genotipos en distintos ambientes. Existen tres patrones distintos de IGA: cambio de magnitud en la diferencia de la respuesta sin cambio de ranking entre genotipos, cambio de ranking sin cambio en la magnitud de las diferencias, y cambio de ranking y de magnitud. El cambio de *ranking* dificulta la selección dado que, el mejor genotipo dependerá del ambiente. La IGA se evidencia por cambios en la correlación del

desempeño de los genotipos y cambios en las varianzas genotípicas entre ambientes (Falconer y Mackay 1996).

En el mejoramiento genético existen distintas formas de abordar la IGA: ignorarla, reducirla o explotarla. Ignorar la presencia de IGA significa seleccionar genotipos por su media en el conjunto de ambientes y es recomendable solo cuando la IGA no involucre cambio de ranking de los genotipos. Reducir la IGA significa agrupar ambientes según similitudes, y mejorar genotipos para cada subgrupo de ambientes (mega-ambiente). Las técnicas más usadas para agrupar ambientes son análisis de clúster y componentes principales. Otra forma de abordar la IGA es explotarla, buscando obtener genotipos específicos según ambientes al igual que en el caso anterior, pero sin reducir la IGA. Se incorporan en esta estrategia los análisis de estabilidad de los genotipos en los distintos ambientes (Crossa et al. 1991; Crossa et al. 2010a) y los modelos que incorporan la IGA en estimación de los valores genéticos de los individuos (Burgueño et al. 2012; Malosetti et al. 2013).

Existen diversos modelos que permiten incluir datos MET, modelando IGA a través de las matrices de varianza-covarianza (Piepho 1998; Malosetti et al. 2004; Burgueño et al. 2012). Entre estos modelos que incorporan la IGA modelando la matriz de varianza-covarianza se destacan los modelos de simetría compuesta, de factores de análisis, y modelos sin estructura, y se pueden representar de la siguiente forma:

$$[5] \quad y = X\beta + Zu + \varepsilon$$

donde y es un vector de datos fenotípicos con las observaciones en cada ambiente; β el vector de efectos ambientales fijos y X la matriz de diseño para estos efectos; u es el efecto aleatorio se asume que $u \sim N(0, G)$ donde G contiene la varianza-covarianzas entre los efectos aleatorios y Z la matriz de diseño de los efectos aleatorios; ε es el residual del modelo de la forma $\varepsilon \sim N(0, R)$. Estos modelos definen distintas estructuras para la matriz G , según se consideren varianzas entre ambientes como homogéneas o heterogéneas (Piepho 1998, Malosetti et al. 2004).

En el modelo de simetría compuesta u del modelo [5] se define como:

$$u = Z_g g + g e$$

donde $Z_g = 1_e \otimes I_g$ es la matriz de diseño para los efectos genotípicos (g); g es el vector aleatorio de efectos genotípicos $g \sim N(0, I_g \sigma_g^2)$; ge es el vector aleatorio de la IGA $ge \sim N(0, I_{ge} \sigma_{ge}^2)$; I_g es una matriz de identidad que indica el relacionamiento entre individuos; I_{ge} una matriz identidad de dimensiones individuos por ambientes; σ_g^2 la varianza genética de los individuos; y σ_{ge}^2 la varianza genética de la IGA. Considerando esta definición, la matriz de varianza-covarianza de los efectos aleatorios (u) será $G = G_e \otimes I_g$ siendo G_e la matriz de varianza-covarianza entre ambientes. Según el modelo de simetría compuesta $G_e = J_e \sigma_g^2 + I_e \sigma_{ge}^2$, siendo J_e una matriz cuadrada de unos, de dimensiones e (número de ambientes). La varianza de un individuo en un ambiente está dada por $\sigma_j^2 = \sigma_g^2 + \sigma_{ge}^2$; la covarianza entre los ambientes j y j' es $\sigma_{jj'} = \sigma_g^2$ para un mismo individuo, y $\sigma_{jj'} = 0$ entre distintos individuos (Smith et al. 2015). En los modelos de simetría compuesta con heterogeneidad el vector aleatorio ge se define como $ge \sim N(0, I_{ge} \otimes D_e)$; siendo D_e una matriz diagonal de dimensión e (número de ambientes) $D_e = \text{diag}\{\sigma_{ge_j}^2\}$ con j el subíndice que indica ambiente, este modelo asume varianzas heterogéneas y requieren que se estime una varianza por ambiente.

Los modelos de factores de análisis (FA) es un modelo de regresión múltiple, del efecto de los genotipos en un ambiente sobre k covariables ambientales con una pendiente por genotipo, en la cual se estima a partir de los datos tanto las pendientes como las covariables (a diferencia de los modelos de regresión múltiple en los cuales las covariables se miden). Considerando el modelo [5], en los modelos FA el vector de efectos aleatorios se definir como: $u_{ij} = \lambda_{1j} f_{1i} + \dots + \lambda_{kj} f_{ki} + \delta_{ij}$, dónde f_{ri} es el valor de los r -ésimo factor hipotético ($r=1$ a k) del i -ésimo genotipo ($i=1$ a g); λ_{rj} es el r -ésimo coeficiente del j -ésimo ambiente ($j=1$ a e); δ_{ij} es el error por falta de ajuste de la regresión múltiple del i -ésimo genotipo en el j -ésimo ambiente. Para continuar con la notación definimos $u = (\Lambda \otimes I_g) f + \delta$; con Λ una matriz de $e \times k$ coeficientes λ y f un vector de largo $g \times k$ de los factores, con $f \sim N(0, I_{gk})$; y $\delta \sim N(0, \psi \otimes I_g)$ dónde ψ es una matriz diagonal de varianza para cada ambiente (conocida como varianza específica). Por lo tanto, para el modelo [5] $u \sim N(0, G_e \otimes I_g)$ donde $G_e = \Lambda \Lambda' + \psi$.

La cantidad de parámetros a estimar por este modelo dependerá de la cantidad de factores hipotéticos que se consideren en la regresión (Piepho 1997; Burgueño et al. 2008; Smith et al. 2015).

Por otro lado, los modelos de varianza-covarianza no estructurados permiten asumir varianzas y covarianzas heterogéneas entre ambientes, por lo que se estima cada uno de los componentes de la matriz G_e . Este tipo de modelos requieren estimar muchos parámetros, por lo que en general son modelos que presentan problemas de convergencia en las estimaciones (Smith et al. 2015).

Burgueño et al. (2012) integraron estos modelos que incorporan la IGA a los modelos G-BLUP usados en SG. De esta forma, la IGA se incorpora en la predicción de los efectos genéticos, donde la matriz de varianza-covarianza de los efectos aleatorios (G) es el producto Kronecker entre la matriz de covarianzas entre ambientes (G_e) y la matriz de relaciones genéticas entre individuos ($G_g = K\sigma_g^2$). Se espera que estos modelos que incluyen IGA en la predicción de nuevas líneas experimentales sean mejores que los modelos que la ignoran al predecir líneas utilizando información de ambientes con IGA. Sin embargo, Burgueño et al. (2012) y Dawson et al. (2013) en sus trabajos no obtuvieron mejoras de la habilidad predictiva de estos modelos, cuando predican nuevas líneas experimentales, respecto a los modelos de medias que ignoran la IGA. Esto se debe a que en estos trabajos, la correlación entre los ambientes fue alta y por ende, hay poca o nula interacción con cambio de *ranking* entre genotipos. Lopez-Cruz et al. (2015) evalúan individuos en diferentes ambientes con presencia de IGA y muestran que el modelo que incluye la interacción es siempre mejor que el modelo de medias en el conjunto de ambientes. Es necesario entonces contar con un mayor número de evaluaciones del comportamiento de estos modelos en grupos de ambientes con distintos niveles de IGA, de forma de definir cuándo el modelo de medias, que es más sencillo, es tan adecuado como el modelo que incluye la IGA.

Una limitante de los modelos anteriormente descritos es que utilizan la correlación ambiental para predecir las líneas en un ambiente dado, por lo que es necesario contar con información del ambiente que se va a predecir. Con el objetivo de superar esta limitante, se desarrollaron modelos que plantean la inclusión de variables ambientales para caracterizar ambientes en los cuales ninguna línea ha sido

evaluada. La inclusión de estas variables permiten incorporar la IGA para incrementar habilidad predictiva de los modelos de SG, sin necesidad de contar con información fenotípica en dichos ambientes (ej. para predecir años futuros; Heslot et al. 2014; Jarquín et al. 2014).

2.5. SELECCIÓN POR CARACTERES MÚLTIPLES

La selección por caracteres múltiples, en general, no es tan efectiva sobre cada carácter en particular como la selección individual sobre el carácter de interés. Sin embargo, en un programa de mejoramiento genético existen varios caracteres objetivo, muchos de los cuáles están genéticamente correlacionados y considerarlos conjuntamente en la selección evita que al seleccionar por un carácter se seleccione negativamente por otro carácter de interés.

2.5.1. Correlación entre caracteres

Dos caracteres fenotípicos pueden estar correlacionados por causas genéticas y/o no-genéticas. La correlación genética es la asociación que nos interesa desde el punto de vista del mejoramiento genético dado que interviene en la respuesta de selección. La correlación no-genética se debe a la acción del ambiente y a la selección. Las causas de correlación genética entre dos caracteres pueden ser ligamiento y pleiotropía. El ligamiento se produce por la asociación entre dos genes cercanos que codifican para los caracteres de interés. Este tipo de correlación es temporal y puede romperse luego de varios ciclos de recombinación. La causa más importante de correlación genética entre dos caracteres es la pleiotropía (Falconer y Mackey 1996), que indica la acción de un gen sobre dos o más caracteres. Este tipo de correlación tiene bases fisiológicas y es estable en el tiempo. Ambos mecanismo permiten obtener valores de correlación positivos o negativos, favorables o desfavorables. La correlación fenotípica está determinada por la correlación genética y la correlación ambiental. En caracteres con altos valores de heredabilidad, en general la correlación genética determina el signo de la correlación fenotípica.

$$r_p = r_a h_x h_y + r_e \sqrt{(1 - h_x^2)(1 - h_y^2)}$$

donde r_p, r_a y r_e son las correlaciones fenotípica, genética aditiva y ambiental, respectivamente, entre los caracteres x e y ; h_x y h_y son las heredabilidades de cada uno de los caracteres.

2.5.2. Respuesta correlacionada a la selección

Si existe correlación entre caracteres existirá una respuesta de selección indirecta en los caracteres correlacionados al carácter por el cual se selecciona. Para poder cuantificar esta respuesta se define el cambio en un carácter (y) dado el cambio en una unidad del carácter correlacionado (x) como:

$$b_{a(xy)} = r_a \frac{\sigma_{a(y)}}{\sigma_{a(x)}}$$

siendo $\sigma_{a(y)}$ y $\sigma_{a(x)}$ los desvíos estándar genéticos aditivos de ambos caracteres x e y ; r_a es la correlación genética entre los caracteres x e y . La respuesta directa de selección en un carácter x (R_x) depende de su heredabilidad (h_x), desvío genético ($\sigma_{a(x)}$) e intensidad de selección (i) para el carácter por el cual se selecciona. La respuesta correlacionada (CR) en y dependerá también de la correlación entre el carácter de interés y el carácter por el cual se selecciona indirectamente:

$$CR_y = b_{a(xy)}R_x = ih_xr_a\sigma_{a(y)}$$

A veces es posible obtener respuestas mayores por selección indirecta que por selección directa lo que dependerá de la heredabilidad de los caracteres y la correlación entre ellos:

$$\frac{CR_y}{R_y} = r_a \frac{h_x}{h_y}$$

2.5.3. Métodos de selección por caracteres múltiples

Cuando se selecciona por caracteres múltiples seleccionando por un carácter a la vez, uno de los métodos más utilizados en el mejoramiento de plantas es la selección en tándem. Una de las formas más comunes de selección en tándem consiste en inicialmente seleccionar plantas por caracteres de alta heredabilidad, que son fáciles de identificar en el campo. En etapas más avanzadas del programa de mejoramiento se seleccionan plantas por caracteres más complejos, de baja heredabilidad. Para la selección en las primeras etapas, se descartan líneas que estén por fuera de un determinado umbral, esto se denomina selección truncada. La selección truncada

afecta menos a la selección de los caracteres en etapas más avanzadas cuando estos están correlacionados con el carácter por el cual se seleccionó en la primera etapa.

Otro enfoque en la selección por múltiples caracteres es el cálculo del mérito genético neto utilizando índices de selección. Los índices de selección son funciones lineales de los múltiples caracteres por los cuales se quiere seleccionar a los cuales se asigna un peso relativo en la selección. Se han construido diferentes índices según la forma de asignar estos pesos a los caracteres. El índice de Smith-Hazel, multiplica cada valor fenotípico de cada carácter por un valor b_i (e.j.: $I = b_1y_1 + b_2y_2 + \dots + b_t y_t$, considerando t caracteres), que se asigna tomando en cuenta varianzas-covarianzas entre caracteres y un coeficiente económico. Por ejemplo, si consideramos dos caracteres, tenemos que $b_1 = (\sigma_{a1}^2 * \sigma_{p1}^2 + \sigma_{a1}\sigma_{a2} * \sigma_{p1}\sigma_{p2}) * e_1$, con σ_{p1}^2 , σ_{a1}^2 la varianza fenotípica y genética aditiva del carácter 1; y $\sigma_{p1}\sigma_{p2}$, $\sigma_{a1}\sigma_{a2}$ la covarianzas fenotípica y genética aditiva entre los caracteres 1 y 2, respectivamente; e_1 es el peso económico asignado al carácter 1 (Smith 1936; Hazel 1943). El índice Smith-Hazel es útil cuando pueden estimarse adecuadamente las varianzas y covarianzas, en caso contrario muchas veces se utiliza el índice base que solo considera pesos económicos (Williams 1962). Cuando los pesos económicos son difíciles de determinar, se utilizan otros índices como por ejemplo el índice multiplicativo, en el cual se asignan valores límites a cada carácter por debajo de los cuales no se seleccionan individuos, lo que es una variación de la selección truncada (Subandi et al. 1973). Otro índice es el de ganancia genética deseada, en el cual se consideran solo las varianzas y covarianzas genéticas aditivas entre caracteres y la ganancia de selección deseada para ese carácter (Pešek y Baker 1969). A veces es deseable obtener mejoras en algunos caracteres y no en otros para lo que puede usarse el índice anterior asignando ganancia cero en los caracteres por los cuales no quiere seleccionarse. Cuando no es posible determinar pesos económicos, ganancias deseadas o límites de selección, puede usarse el índice retrospectivo que utiliza el peso relativo de los caracteres que se ha asignado históricamente en el programa de mejoramiento (Bernardo 1991). Este índice es similar al índice Smith-Hazel, pero utiliza únicamente varianzas y covarianzas fenotípicas y el diferencial de selección en vez de los pesos económico. El diferencial

se calcula como la media de los individuos seleccionados menos la media de la población.

Los índices de selección se han desarrollado previamente para selección fenotípica y este enfoque se ha comenzado a extender incorporando la información genotípica, primero en el contexto de la selección asistida por marcadores (Lande y Thompson 1990) y actualmente para su inclusión en SG (Dekkers 2007; Ceron-Rojas et al. 2015). Los índices de SG combinan linealmente los valores genéticos de los individuos calculados para cada carácter con el objetivo de predecir el mérito genético neto.

2.5.4. Predicciones de caracteres múltiples en selección genómica

La SG presenta un desafío, ya que implica seleccionar tempranamente líneas por muchas de sus características para avanzarlas generacionalmente lo más rápido posible. Esto implica seleccionar en una única etapa del mejoramiento por varias características al mismo tiempo. Con este objetivo se ha investigado la inclusión de caracteres múltiples en los modelos de predicción genómica.

Los modelos multi-carácter aumentan la información fenotípica disponible en los modelos de SG (ej. GBLUP), incluyendo información de varianzas y covarianzas de varios caracteres que responde a los mismos genes o a genes ligados entre sí (Calus y Veerkamp 2011; Guo et al. 2014):

$$y = X\beta + Zg + \varepsilon$$

donde y es un vector de datos fenotípicos del conjunto de caracteres; β el vector de efectos fijos y X la matriz de diseño para estos efectos; g es el efecto aleatorio de los genotipos para cada carácter y se asume que $g \sim N(0, \rho \otimes K)$; ρ es la matriz de varianza-covarianza entre caracteres, K es la matriz de relaciones genéticas estimada con marcadores moleculares; Z la matriz de diseño de estos efectos aleatorios; ε es el residual del modelo de la forma $\varepsilon \sim N(0, R)$, $R = \text{diag}\{\sigma_{e_t}^2\}$ con $\sigma_{e_t}^2$ la varianza residual para cada carácter.

Los modelos multi-carácter no han sido mejores que los de un solo carácter, dado que solo presentaron mayor habilidad predictiva en situaciones simuladas, donde un carácter de muy baja heredabilidad se predice utilizando caracteres muy correlacionados y de alta heredabilidad (Jia y Jannink 2012). Sin embargo, se han

demostrado su utilidad para optimizar los recursos invertidos en un programa de mejoramiento. Se han logrado obtener mayores precisiones en la predicción de caracteres costosos o difíciles de medir, utilizando en la predicción caracteres correlacionados que sean más económicos o de fácil medición (Calus y Veerkamp 2011; Jia y Jannink 2012). Actualmente, el desarrollo de métodos de fenotipado permite evaluar muchos caracteres con alta precisión en muchas líneas. A través de la aplicación de los modelos multi-carácter es posible incrementar la precisión de las predicciones de caracteres difíciles de evaluar (ej. rendimiento) utilizando caracteres correlacionados de más fácil medición (Rutkoski et al. 2015; Hayes et al. 2017; Sun et al. 2017). Estos modelos multi-carácter son una herramienta poderosa para ser utilizada en la optimización de la asignación de recursos dentro de un programa de mejoramiento, que aún no se ha explorado completamente.

2.6. PREDICCIÓN DE CRUZAMIENTOS

La selección de los padres a cruzar entre las líneas elite de un programa es un paso clave del mejoramiento genético, en el cual existen muchas combinaciones posibles de padres que darán lugar a muchos cruzamientos (Bernardo 2003). Seleccionar con mayor información los mejores cruzamientos evitaría que muchos de estos cruzamientos sean posteriormente descartados en su totalidad. Con una mayor selectividad en la elección de cruzamientos, se podrá evaluar un mayor número de líneas por cruce, dado que los recursos de un programa de mejoramiento son siempre limitados (Witcombe et al. 2013).

En la selección de los cruzamientos se debe tomar en cuenta el desempeño de los padres a seleccionar, debido a que la media de la performance de ambos padres es un buen predictor de la media de la progenie del cruzamiento de ambos (Bernardo 2014). Si se evalúan progenies con igual valor medio en su desempeño, parecería evidente que vamos a preferir seleccionar el grupo de líneas que presente mayor varianza, ya que aumentamos la probabilidad de obtener individuos en la progenie con un desempeño superior. Por lo tanto es importante considerar el rol de la varianza genética en la selección (Souza y Sorrells 1991).

Existen trabajos que intentan predecir la varianza genética de la progenie tomando en cuenta diferencias morfológicas y genéticas entre ambos padres (Souza y Sorrells 1991; Melchinger et al. 1998; Utz et al. 2001; Bertan et al. 2007; Hung et al. 2012; Casassola y Brammer 2013). Hasta ahora sin embargo, no se ha encontrado un buen método para predecir la varianza de la progenie. Sin embargo, dada la actual disponibilidad de marcadores a lo largo de todo el genoma, Endelman (2011) y Bernardo (2014) proponen estimar esta varianza considerando el efecto de los marcadores que son diferentes entre ambos padres. Mohammadi et al. (2015) y Tiede et al. (2015) proponen a su vez, la estimación de la varianza de la progenie considerando también el desequilibrio de ligamiento presente entre los marcadores. Estos métodos de estimación de la varianza requieren su comparación y evaluación con datos experimentales que permitan conocer la precisión en la estimación de las varianzas genéticas.

3. MODELADO DE LA INTERACCIÓN GENOTIPO POR AMBIENTE

3.1. RESUMEN DEL CAPÍTULO

La SG implica predecir el comportamiento de nuevos individuos utilizando información de una población de entrenamiento. Dicha población cuenta con información genotípica y fenotípica que es utilizada por el modelo para predecir el comportamiento de los nuevos individuos. Cuanto más detallada sea la información fenotípica, más precisa serán las predicciones del modelo. Por lo tanto, para entrenar el modelo de predicción se utiliza información de múltiples años y localidades. El uso de este tipo de información requiere analizar la presencia de IGA para estudiar cómo utilizarla en el contexto de SG. En este trabajo se comparó la habilidad predictiva de un modelo de medias que ignora la IGA ($GBLUP_M$) y un modelo que la incluye ($GBLUP_{gxe}$) en la predicción de individuos desconocidos e individuos parcialmente fenotipados entre grupos de ambientes con distintos niveles de IGA. Para esto, se utilizaron 1477 líneas avanzadas del programa de mejoramiento de trigo genotipadas con 81999 marcadores SNP, utilizando la técnica de GBS. La información fenotípica de rendimiento de grano se recolectó en 35 ambientes (combinaciones de 5 años y 5 localidades), en los cuales se sembraron distintos tipos de ensayo según la etapa del programa (preliminar, avanzada y de elite) y según el ciclo de maduración del cultivo (corto, intermedio y largo). Al estudiar la IGA, se identificaron tres mega-ambientes separados principalmente por IGA significativa entre años, donde principalmente los mega-ambientes 1 y 2 reunieron los años de mejores rendimientos (2010, 2011 y 2013) y el mega-ambiente 3 los de menor rendimiento (2012 y 2014). En la predicción de nuevos individuos, el modelo $GBLUP_{gxe}$ fue mejor que $GBLUP_M$ para ambientes con mayor IGA. Sin embargo, ambos modelos fueron similares cuando se redujo la IGA (localidades y mega-ambientes). Por otro lado, en la predicción de individuos parcialmente fenotipados en otros ambientes, no se encontró un mejor modelo al predecir distintos grupos de ambientes con distintos niveles de IGA.

Modeling Genotype \times Environment Interaction for Genomic Selection with Unbalanced Data from a Wheat Breeding Program

Bettina Lado, Pablo González Barrios, Martín Quincke, Paula Silva, and Lucía Gutiérrez*

ABSTRACT

Genomic selection (GS) has successfully been used in plant breeding to improve selection efficiency and reduce breeding time and cost. However, there is not a clear strategy on how to incorporate genotype \times environment interaction (GEI) to GS models. Increased prediction accuracy could be achieved using mixed models to exploit GEI by borrowing information from other environments. The objective of this work was to compare strategies to exploit GEI in GS using mixed models. Specifically, we compared strategies to predict new genotypes by borrowing information from other environments modeling the correlation matrix across environments and to design sets of environments aiming for low GEI to predict genomic performance in new environments. We evaluated 1477 advanced wheat (*Triticum aestivum* L.) lines for yield in 35 location-year combinations genotyped with genotyping-by-sequencing (GBS). Mixed models were used to obtain either overall or by-environment predictions for different sets of environments. Overall accuracy was high (0.5). Borrowing information from relatives evaluated in multiple environments and modeling the correlation matrix across environments was the best strategy to predict new genotypes. On the other hand, the best strategy for predicting the performance of genotypes in new environments was either to predict across locations for single years or to predict within defined mega-environments (MEs) for any year or location. In summary, higher predictive ability was obtained by characterizing and by modeling GEI in the GS context.

B. Lado, P. González Barrios, and L. Gutiérrez, Dep. of Statistics, College of Agriculture, Universidad de la República, Garzón 780, Montevideo, Uruguay; L. Gutiérrez, Dep. of Agronomy, Univ. of Wisconsin–Madison, 1575 Linden Dr., Madison, WI 53706; M. Quincke and P. Silva, Programa Nacional de Investigación Cultivos de Secano, Instituto Nacional de Investigación Agropecuaria, Est. Exp. La Estanzuela, Colonia 70000, Uruguay. Received 2 Apr. 2015. Accepted 14 Sept. 2015. *Corresponding author (gutierrezcha@wisc.edu).

Abbreviations: ALL, all location-year combinations; AYT, advanced yield trial; BLUE, best linear unbiased estimation; CV, cross-validation; DOL, location Dolores; DZ, location Durazno; GBLUP_{g \times e}, by-environment genomic predictions; GBLUP_M, overall genomic predictions for the mean of the environments; GBS, genotyping-by-sequencing; GEI, genotype \times environment interaction; GGE, genotypic main effect and genotype \times environment interaction matrix; GS, genomic selection; INIA, National Agricultural Research Institute; LE, location La Estanzuela; ME, mega-environment; MET, multi-environment trial; PYT, preliminary yield trial; R2, location Ruta2; WBP, Wheat Breeding Program; YET, elite yield trial; YOU, location Young.

GENOMIC SELECTION has proven highly successful, and plant breeding programs are implementing GS strategies as part of their breeding routines (Heffner et al., 2010; Lorenz et al., 2012; Hayes et al., 2013; Rutkoski et al., 2015). Genomic selection was first developed in animal breeding (Meuwissen et al., 2001) and has since been transferred into plant breeding (Heffner et al., 2009). While most of the GS studies have focused on evaluating GS model performance (Bennewitz et al., 2009; Solberg et al., 2009; Heslot et al., 2012; de los Campos et al., 2013) and optimizing the training population in terms of population size (Lorenzana and Bernardo, 2009; Asoro et al., 2011; Lorenz et al., 2012), number of markers (Lorenzana and Bernardo, 2009; Asoro et al., 2011; Heffner et al., 2011), and its structure (Asoro

Published in Crop Sci. 56:1–15 (2016).
doi: 10.2135/cropsci2015.04.0207

© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA
This is an open access article distributed under the CC BY-NC-ND license
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

et al., 2011; Crossa et al., 2013; Isidro et al., 2014); there are few reports demonstrating results from implementing GS in the full context of a plant breeding program. One of the key differences between animal and plant breeding is that in the latter, GEI is large (Cooper and Delacy, 1994; Mathews et al., 2008). Multi-environment trials (MET) are routinely conducted in plant breeding to capture and account for GEI. However, these trials are highly unbalanced because of poor lines being removed from evaluation while promising lines are further evaluated (Bernardo, 2010). To implement GS strategies in plant breeding, GEI needs to be accounted for, modeled, and predicted (Resende et al., 2011; Burgueño et al., 2012; Heslot et al., 2014; Lopez-Cruz et al., 2015).

Several strategies have been proposed to deal with GEI: ignore, reduce, or exploit it (Eisemann et al., 1990; DeLacy et al., 1996). Ignoring or reducing GEI is possible when the proportion of the total phenotypic variance as a result of GEI is low (DeLacy et al., 1996) or when selection for broad adaptation is conducted (Lal, 2013; Lal, 2014). However, the most powerful strategy is to exploit GEI to either develop locally adapted material (Mathews et al., 2008) or to use GEI to better characterize the genotypes (Dreccer et al., 2007; Mathews et al., 2008; Cooper et al., 2014). Genotype \times environment interaction was first characterized with linear-bilinear models using the singular value decomposition of either the GEI matrix (i.e., additive main effect and multiplicative interaction, AMMI; Gollob, 1968; Mandel, 1969; Gauch and Zobel, 1988; van Eeuwijk, 1995; Gauch, 2006; Gauch et al., 2008; Crossa, 2012); the genotypic main effect and GEI matrix (GGE; Yan et al., 2000, 2007; Yan and Tinker, 2006); or site regression (Crossa and Cornelius, 1997). Linear-bilinear models could be graphically represented through biplots (Gabriel, 1971; Yan and Tinker, 2006) or enhanced biplots (Gauch, 2006), where genotypes and environments are plotted in a single graph. Later, mixed models were introduced to allow more flexibility and to model different correlation structures among environments (Piepho, 1998; Burgueño et al., 2008; Kelly et al., 2009; Cullis et al., 2010), increasing statistical power (Wang and Schmidt, 2002). A number of correlation structures can be modeled with mixed models, from a single correlation among environments (i.e., compound symmetry models; Piepho, 2005), heterogeneous variances (Piepho, 1998; Malosetti et al., 2004; Burgueño et al., 2008; Cullis et al., 2010), up to a unique correlation for each pair of environments (i.e., unstructured variance-covariance; Burgueño et al., 2008; Malosetti et al., 2013). Finally, MEs were defined as groups of environments with minimal GEI within (Braun et al., 1996). Both linear-bilinear and mixed models are powerful tools to characterize GEI and have successfully been used in plant breeding.

Breeding programs produce large MET datasets that are highly unbalanced with heterogeneous quality because of the degree of replication in different trials (Bernardo, 2010). Early generations usually consist of a large number of genotypes evaluated in a few environments with few replications. Advanced trials, on the other hand, consist of few genotypes evaluated in a large number of environments with more replications (Bernardo, 2010; Lorenz et al., 2012). This creates a high level of unbalance in the data sets that can be properly managed with mixed models using restricted maximum-likelihood estimates (Patterson and Thompson, 1971). Mixed models can also model individual-plot variability to account for variability in data quality from MET (Malosetti et al., 2004; Mathews et al., 2008). In summary, mixed models are flexible enough to model MET data with high levels of unbalance, with correlation between environments and genotypes, and with variability in data quality from multiple trials.

Genotype \times environment interaction has been exploited in genomic studies to map quantitative trait loci in balanced populations (Piepho, 2000; Malosetti et al., 2004; van Eeuwijk et al., 2005; Boer et al., 2007; Mathews et al., 2008) and genome-wide association studies (Gutiérrez et al., 2015) using mixed models. Additionally, GS studies have included GEI information by performing overall predictions across environments (Heffner et al., 2011; Resende et al., 2011), within environments or groups of environments (Burgueño et al., 2012; Ly et al., 2013; Dawson et al., 2013; Heslot et al., 2014), or using marker-by-environment predictions (Jarquín et al., 2014; Lopez-Cruz et al., 2015). While mixed models have successfully been used to model GEI in GS, the best strategy to predict new genotypes or new environments is not clear. The aim of this work was to improve the predictive ability of the models comparing strategies to exploit GEI in GS to predict new genotypes and new environments by evaluating different groups of environments and modeling the variance-covariance matrix across environments in a large, highly unbalanced, historical dataset from a wheat breeding program. We specifically aimed to answer the following questions: Could we improve the predictive ability by modeling the variance-covariance matrix across environments and predicting the performance of new genotypes (i.e., unphenotyped genotypes)? Could we design groups of environments where overall predictions outperformed environment-specific predictions? Could we borrow information from some environments to predict the performance in new environments (i.e., untested environments)?

MATERIALS AND METHODS

Plant Material

A total of 1477 advanced inbred lines from the National Agricultural Research Institute (INIA) Wheat Breeding Program (WBP) were extensively phenotyped and also genotyped to be

Table 1. Description of the wheat breeding program trials indicating the number of genotypes in each trial or group of trials, mean yield (kg ha⁻¹) per trial, and standard deviation for 1477 genotypes evaluated in 35 environments in preliminary (PYT), advanced (AYT), and elite (EYT) yield trials grouped by cycle (S, short; I, intermediate; L, long).

Year	Location [†]	EYT _{SI}	EYT _L	AYT _S	AYT _I	AYT _L	PYT _S	PYT _I	PYT _L	No. genotypes	Yield	SD
2010	DOL	25			64	36				111	5358	455
	LE1	25	16		64	36	300	197	326	617	4835	1284
	LE2	25	16							41	5566	1265
	LE3	25	16							41	4875	1151
	LE4	25	16							41	4972	1455
	R2	25								25	5336	526
2011	YOU	25	16		64	36				117	5366	681
	DOL	30	16	81	49	42				159	4028	857
	DZ	30	16							44	5758	688
	LE1	30	16	81	49	42	239	128	264	957	4652	1416
	LE2	30	16							44	5560	631
	LE3	30	16							44	5439	719
2012	LE4	30	16							44	3182	1092
	R2	30	16							44	6476	504
	YOU	30	16	81	49	42				110	4546	649
	DOL	30	20	49	49	42				138	2624	604
	LE1	30	20	49	49	42				138	2482	1138
	LE2	30	20							48	3342	1029
2013	LE3	30	20							48	3437	1050
	LE4	30	20							48	4042	1048
	R2	30	20							48	3323	1050
	YOU	30	20	49	49	42				138	1920	734
	DOL	30	20							50	5782	470
	LE1	30	20				233	199	149	631	6737	1191
2014	LE2	30	20							50	7819	672
	LE3	30	20							50	5667	867
	LE4	30	20							50	5055	862
	R2	30	20							50	7714	655
	YOU	30	20							50	4444	694
	DOL	30	16	64	49	36	132	69	132	30	3538	728
2014	LE1	30	16	64	49	36				528	4341	1125
	LE2	30	16							46	4581	1034
	LE3	30	16							45	4166	1063
	R2	30	16	64	49	36				195	4305	614
	YOU	30	16	64	49	36				195	3910	814

[†] DOL, Dolores; DZ, Durazno; LE, La Estanzuela; R2, Ruta2; YOU, Young.

used as a training set for genomic prediction models. These genotypes were chosen to represent the advanced inbred lines at a given time in the breeding program and consisted of all the lines from the preliminary yield trials (F₇; PYT) from 2010, 2011, and 2013 as well as the advanced (F₈; AYT) and elite (F₉; EYT) yield trials from 2010.

Phenotypic Data

Historic data from the INIA-WBP was used to evaluate yield of genotypes. Phenotypic evaluation from five locations in Uruguay evaluated in 5 yr, including one location with four sowing dates, were used to define a total of 35 environments (Table 1). Dolores (DOL; 33°50' S, 58°14' W, 15 m asl), Durazno (DZ; 33°33' S, 56°31' W, 91 m asl), La Estanzuela (LE; 34° 20' S, 57° 42' W, 81 m asl), Young (YOU; 32°76' S, 57°57' W, 85 m asl), and Ruta2 (R2; 33°45' S, 57°90' W, 95 m asl) were the five locations used to evaluate the genotypes. Four sowing dates

were evaluated in LE (LE1, LE2, LE3, and LE4). The evaluation years were 2010 to 2014. However, the data is typical for a breeding program, as not all genotypes were evaluated in all the environments, and therefore, the data is highly unbalanced in years and locations as a result of promising genotypes being evaluated in multiple environments and poor genotypes being evaluated in a single environment (Table 1; Fig. 1).

Within each year, genotypes were evaluated in PYT, AYT, or EYT. Preliminary yield trials consisted of approximately 600 genotypes each year evaluated in a sequence of trials. Three to twelve PYTs were conducted every year with 50 to 90 genotypes in each trial. Genotypes were grouped in trials by heading date (i.e., short, PYT_S; intermediate, PYT_I; or long, PYT_L cycle). Each trial consisted of a resolvable incomplete block design with two replications where 6 to 10 incomplete blocks nested within each complete replications were used. Only location LE1 was used for PYT (Table 1).

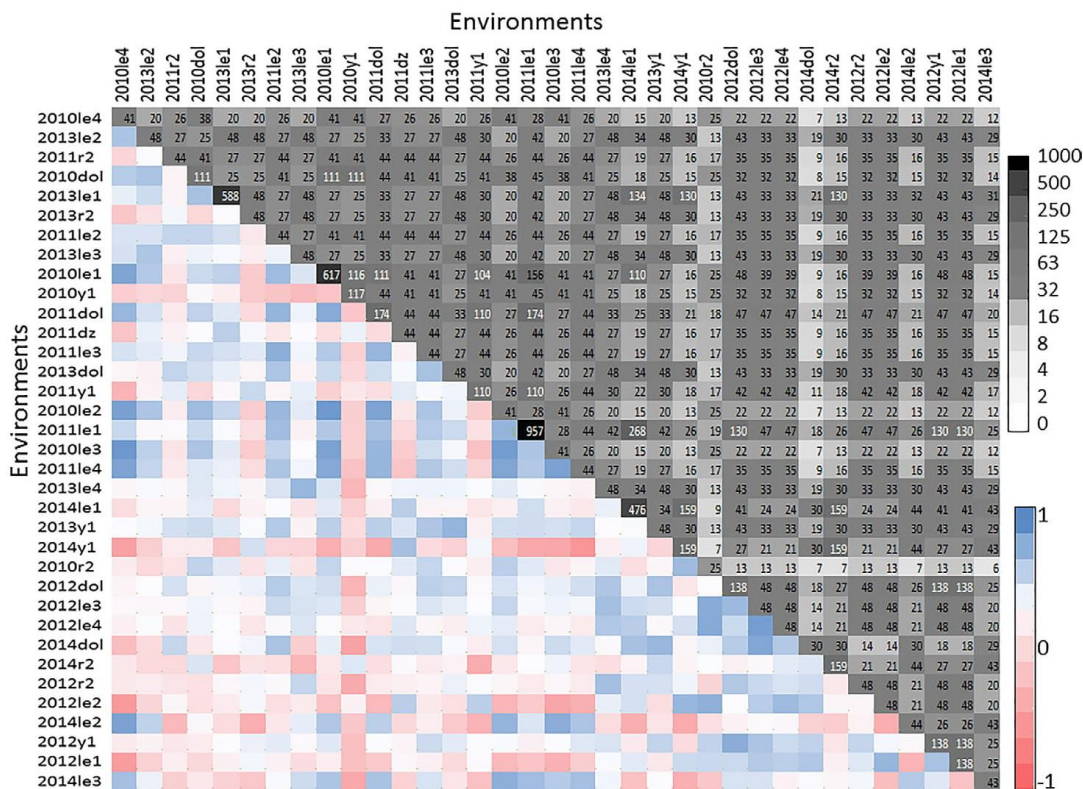


Figure 1. Characterization of genetic correlation across environments and degree of unbalance of 1477 wheat genotypes evaluated in 35 environments. Lower diagonal heat plot indicates genetic correlations among pairs of environments, while upper diagonal indicates the number of shared genotypes evaluated in each pair of environments. Environments are labeled by their year (i.e., 2010–2014) followed by their location (i.e., Dolores, Duranzo, La Estanzuela [LE1–LE4], Ruta2, and Young).

Advanced yield trials consisted of approximately 150 genotypes each year evaluated in a series of trials. Three AYT_s were conducted every year with 35 to 80 genotypes in each trial with genotypes grouped in trials by heading date (i.e., short, AYT_s; intermediate, AYT_i; or long, AYT_L cycle). Each trial consisted of a resolvable incomplete block design with three replications where six to nine incomplete blocks nested within each complete replication were used.

Elite yield trials consisted of approximately 50 genotypes each year evaluated in a series of trials. Three EYT_s were conducted every year with 16 to 30 genotypes in each trial with genotypes grouped by heading date (i.e., short and intermediate, EYT_{SP}, or long, EYT_L cycle). Each trial consisted of a resolvable incomplete block design with four replications where four to five incomplete blocks nested within each complete replication were used. All five locations were used for EYT_s (Table 1). A total of 52 historic as well as commercial varieties were used as checks in the trials having each trial between eight and 20 checks at a time.

Genotypic Data

The 1477 wheat lines were genotyped by GBS (Elshire et al. [2011] modified by Poland et al. [2012a] for wheat). The SNP identifications were conducted using the Tassel pipeline (Glau-bitz et al. [2014] as in Lado et al. [2013]). Briefly, markers with a minor allele frequency smaller than 0.01 and more than 80% of missing data were discarded. We identified 81,999 SNPs. Marker-data imputation was conducted using the realized relationship matrix through the multivariate normal expectation maximization method using the rrBLUP package (Endelman, 2012) from R software (R Development Core Team, 2015).

Phenotypic Data Analysis

Phenotypic best linear unbiased estimations (BLUES) were obtained for all genotypes present in each trial. Field analysis was conducted according to experimental design. Since PYTs consisted of a series of smaller α -design trials connected through common checks, the following model was used to estimate genotypic means for each heading date group in each environment:

$$y_{ijkl} = \mu + g_i + t_j + r_{k(j)} + b_{l(jk)} + \varepsilon_{ijkl} \quad [1],$$

where μ is the overall mean, g_i is the fixed effect of the i th genotype, t_j is the random effect of the j th trial, $r_{k(j)}$ is the fixed effect of the k th replicate nested within the j th trial, $b_{l(jk)}$ is the random effect of the l th incomplete block nested within the j th trial and the k th replicate, and ε_{ijkl} is the residual error for the i th genotype in the l th incomplete block within the k th replicate in the j th trial with t_j , $b_{l(jk)}$ and ε_{ijkl} as random variables being $t_j \sim N(0, \sigma_t^2)$, $b_{l(jk)} \sim N(0, \sigma_b^2)$ and $\varepsilon_{ijkl} \sim N(0, \sigma_e^2)$.

The advanced and elite yield trials (AYT and EYT) consisted of α -designs, and therefore the following model was used to estimate genotypic means for each heading date group in each environment:

$$y_{ikl} = \mu + g_i + r_k + b_{l(k)} + \varepsilon_{ikl} \quad [2],$$

where μ , g_i , r_k , $b_{l(k)}$, and ε_{ikl} were defined as in Eq. [1], with $b_{l(k)}$ and ε_{ikl} as random variables being $b_{l(k)} \sim N(0, \sigma_b^2)$ and $\varepsilon_{ikl} \sim N(0, \sigma_e^2)$. Best linear unbiased estimates were estimated for each genotype in each heading date group and each environment.

Broad-sense heritability was estimated for each trial. We estimated variance component using restricted maximum likelihood method estimates from Eq. [1] or [2] but with genotypes as random effect. All of the analyses were conducted in R software (R Development Core Team, 2015) with package nlme (Pinheiro et al., 2013).

Characterization of Genotype \times Environment Interaction

Several strategies were used to characterize GEI including variance component estimation, correlation between environments, AMMI and GGE models, and graphical representation through augmented biplots.

The magnitude of GEI for yield was first characterized with variance components for all the years and locations. A random effect model using the PROC MIXED procedure of SAS (SAS Institute, 2005) was used to estimate variance components for each yield trial (i.e., PYT, AYT, and EYT). Second, Pearson correlations between all pairs of environments using all the shared lines between each pair of environments were estimated.

Multiplicative models such as AMMI (Gauch and Zobel, 1988; Zobel et al., 1988) and GGE (Yan et al., 2000) were used to further characterize GEI. Since most of the GEI structure is given by the EYT that are evaluated in multiple locations, a subset consisting of 60 elite genotypes were chosen to construct both AMMI and GGE models. These genotypes were chosen to have at least 50% coverage in the genotype \times environment table of means and because the GEI structure they create is similar to that of the complete set of genotypes that have a 7% coverage (Fig. 1; Supplemental Fig. S1). A mixed model with the realized additive relationship matrix to model covariance among genotypes and the estimated correlation matrix across environments was fitted to predict genotypic performance of all genotypes in all the environments before fitting AMMI and GGE models (Henderson, 1984; Bernardo, 2010). The genetic correlation matrix between environments was estimated using the spectral decomposition method described by Rebonato and Jäkel (1999) and subsequently used as the variance-covariance matrix. The mixed model for genomic prediction was fit with

the rrBLUP package (Endelman, 2012) using the Kronecker product between the additive relationship matrix among genotypes and the estimated variance-covariance matrix among environments. This model assumes equal genotypic relationship matrix in all environments (for further description of this model, see the following section). The AMMI1 to AMMI10 and GGE1 to GGE10 were fitted in SAS statistical software (SAS Institute, 2005) and the Gollob's (1968) statistical test was used to choose the number of significant axis from both AMMI and GGE model families. Environments were grouped in MEs with the best AMMI and GGE models. Additionally, we compared the MEs formed with the best models to those of the simplest models: GGE2 and AMMI2. Finally, since the MEs created by the different strategies were similar, a GGE2 augmented biplot representation was used to visually represent the GEI structure using the R statistical software (R Development Core Team, 2015) with the GGEBiplotGUI package (Frutos et al., 2014).

Genomic Predictions and Prediction Accuracy

To compare strategies to deal with GEI, we partitioned the data in 13 sets of environments: by year (i.e., all locations for 2010, 2011, 2012, 2013, or 2014), by location (i.e., all years for DOL, LE1, R2, or YOU), by ME (i.e., all location-year combinations within each ME1, ME2, and ME3), and all environments (all location-year combinations [ALL]). We used two general approaches to make genomic predictions: overall predictions for the mean performance within each set of environments (GBLUP_M, i.e., predictions of genomic estimated breeding values [GEBVs] for the mean performance of genotypes in each set of environments: years, locations, ME, or ALL), and predictions by environment (GBLUP_{g \times e}, i.e., predictions of GEBVs for each environment within sets created by years, locations, ME, or ALL). We used the following general model structure to make predictions:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{g} + \boldsymbol{\varepsilon} \quad [3]$$

For GBLUP_M: $\mathbf{y}_{N \times 1}$ is the vector of mean yield for each genotype in the set of 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), $\mathbf{1}_{N \times 1}$ is a vector of ones of length N ; μ is the overall mean; $\mathbf{g}_{N \times 1}$ is a random vector of genotypic predictors with $\mathbf{g} \sim N(0, \mathbf{A}_{N \times N} \sigma_g^2)$, where \mathbf{A} is the realized additive relationship matrix; $\mathbf{Z}_{N \times N}$ is the corresponding incidence matrix; and $\boldsymbol{\varepsilon}$ is the residual errors vector with $\boldsymbol{\varepsilon} \sim N(0, \mathbf{R}_{N \times N})$ where \mathbf{R} is a diagonal matrix with elements as the reciprocals of the variances of the adjusted means for each genotype from the phenotypic analysis stage to account for heterogeneity in mean estimate precision following Rodrigues et al. (2014).

In GBLUP_{g \times e}: $\mathbf{y}_{n \times 1}$ is the mean vector of yield for each genotype in each environment (i.e. BLUEs from field experiments) with length n (n = number of genotypes [N] by number of environment [k]; $N \times k$); $\mathbf{1}_{n \times 1}$ is a vector of ones of length n ; μ is the overall mean; $\mathbf{g}_{n \times 1}$ is a random vector of genotype \times environment predictors with $\mathbf{g} \sim N[0, (\mathbf{A}_{N \times N} \otimes \boldsymbol{\rho}_{k \times k}) \sigma_g^2]$, where $\boldsymbol{\rho}_{k \times k}$ is the estimated correlation matrix among environments and k is the

number of environments; $\mathbf{Z}_{n \times n}$ is the corresponding incidence matrix; and $\boldsymbol{\varepsilon}$ is the residual errors vector with $\boldsymbol{\varepsilon} \sim N(0, \mathbf{R}_{n \times n})$, where \mathbf{R} is a diagonal matrix with elements as the reciprocals of the variances of the adjusted means for each genotype from the phenotypic analysis stage to account for heterogeneity in mean estimate precision following Rodrigues et al. (2014).

Accuracy of GEBVs was estimated as the correlation between predicted and observed genotypic values using cross-validation (CV, Legarra et al., 2008). A k -fold CV (CV_1) 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 set, 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 CV with $k = 7$ were performed and the correlation between GEBV and observed genotypic values was used to estimate the accuracy of GEBV.

Predictions of New Genotypes (Unphenotyped Genotypes)

To predict new genotypes we compared the $GBLUP_M$ with the $GBLUP_{g \times e}$ models. To make unbiased comparisons, the predictive ability of the $GBLUP_M$ model was estimated with the sevenfold CV (CV_1). In the case of the $GBLUP_{g \times e}$ model, the sevenfold CV was modified (CV_2) to predict breeding values for a genotype without any phenotypic information per se from the same genotype in other environments. Briefly, we split the genotype \times environment data into seven subsets. However, to ensure fairness when predicting new genotypes, no information from a given genotype was used to predict its own performance. Therefore, yield data from multiple environments for a single genotype were grouped in the same k th subset. Additionally, sets were constructed to have the same amount of information (i.e., data points) and not necessarily the same number of genotypes. Therefore, the number of genotypes in each subset is slightly different. We also used 100 replications of the CV to estimate the correlation between GEBV and observed genotypic values.

Finally, because the sets of environments include different number of genotypes (i.e., population size), we compared the predictive ability of $GBLUP_M$ and $GBLUP_{g \times e}$ models for all sets of environments using a series of random samples with different sizes for all sets (i.e., $N = 99$, $N = 200$, $N = 500$, and $N = 900$). We compared the predictive ability using the sevenfold CV described earlier (CV_1 and CV_2).

Predictions of New Environments (Unphenotyped Environments)

Different sets of environments were used to predict genotypes in new environments (unphenotyped environments) using both GBLUP models. First, we considered environments grouped by year and we predicted the genomic breeding values for the tested genotypes in each location within a year using all the remaining locations for that year. This way, the training set consisted of all genotypes in all the locations less one, and the testing set consisted of the same genotypes in the remaining location. We followed this procedure to predict genomic breeding values for all genotypes in all locations one at a time. The accuracy of the

genomic prediction was estimated as the correlation between genomic estimated breeding values for a specific location and the observed yield performance of genotypes for that location. Second, we used the same principle but for environments grouped by location. We used all years less one for a specific location as the training set while we used the remaining year as the testing set. Finally, we used the same principle but for environments grouped in MEs. We trained the model with all environments (i.e., combination of locations and years) in a ME less one, and we tested it with the remaining environment for that ME.

RESULTS

Phenotypic Data Analysis

The dataset is highly unbalanced having pairs of environments sharing from six to 268 genotypes (Fig. 1). The LE1 location included PYT, AYT, and EYT, it is therefore the environment with the largest number of genotypes (Table 1; Fig. 1). The other locations with the largest number of genotypes were YOU in all years, DOL in 2010 to 2013, and R2 in 2014.

The broad-sense heritability estimated on a line-mean basis was high for most trials (i.e., >0.65 with an average of 0.82; Fig. 2), and only four trials had low heritability (i.e., 0.26 and 0.54 in EYT in R2 in 2013 and 0.42 and 0.47 in LE1 in 2010; Fig. 2).

Environmental Characterization of Genotype \times Environment Interaction

Genotype \times environment interaction contributed a large proportion of the total variance, being 7.6, 16.9, and 15.5% of the total variance in the PYT, AYT, and EYT, respectively (Table 2). Specifically, GEI explained much more than the genotypic variance in AYT and EYT, where several years were evaluated. The year component was large in all cases and GEI among years is larger than GEI among locations (Table 2).

Eight components were significantly different from zero in both AMMI and GGE models (Table 3). The best GGE model family (i.e., GGE8) identified three MEs that were the same as the simplest model (i.e., GGE2, data not shown); ME1 (with 11 location-year combinations) and ME2 (with 10 location-year combinations) included all locations for 2010, 2011, and 2013 while ME3 (with 14 location-year combinations) included all locations for 2012 and 2014 (Fig. 3). In GGE models, the first two components explained 74% and the first eight explained 97% of total variance (genotype plus GEI variance). The best one of the AMMI model family (i.e., AMMI8) grouped environments into two MEs collapsing ME1 and ME2 into a single group and discriminating years 2012 and 2014 from the remaining years (data not shown). In the AMMI models, the first two components explained 74%, and the first eight components explained 96% of the total variance (GEI variance). The environments with the largest discriminative ability were LE4 in 2011, LE2 and YOU in 2013, LE1 and YOU

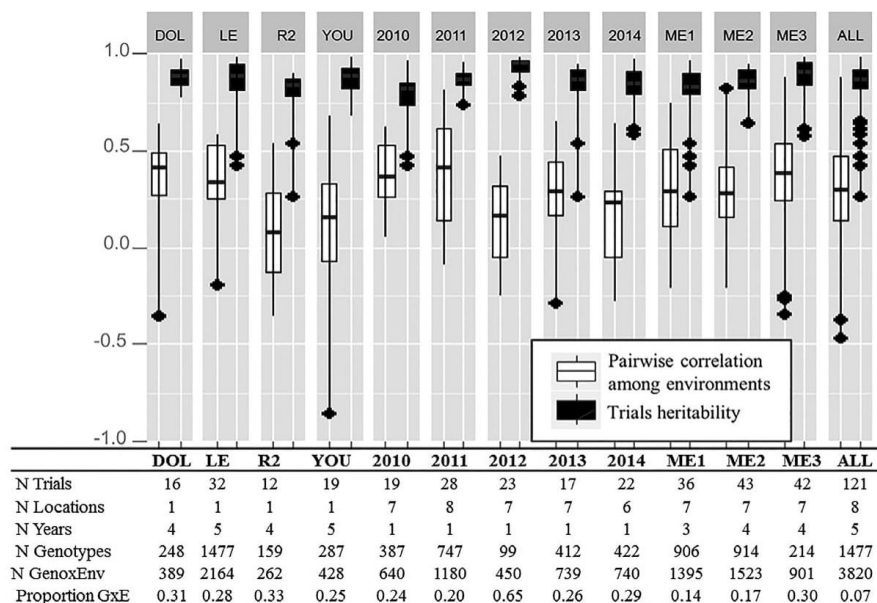


Figure 2. Pairwise correlations among environments and heritability for all trials (PYT_S, PYT_L, PYT_L, AYT_S, AYT_L, AYT_L, EYT_S, and EYT_L) within sets of environments. The number of trials, years, locations, and the proportion of the genotype × environment interaction for each set of environments is indicated. The 35 environments are grouped by either years (i.e., 2010–2014), locations (i.e., Dolores [DOL], La Estanzuela [LE1], Ruta2 [R2], Young [YOU]), mega-environment (ME1–ME3), or all together (ALL).

Table 2. Yield variance component estimation and proportion of the total variance for preliminary (PYT), advanced (AYT), and elite (EYT) yield trial for 1477 wheat genotypes evaluated in 35 environments that are a combination of locations and years.

Variance Component	PYT		AYT		EYT	
	Yield	Proportion	Yield	Proportion	Yield	Proportion
	kg ha ⁻¹	%	kg ha ⁻¹	%	kg ha ⁻¹	%
Location	–	–	100576	3.2	307877	10.2
Year	1640866	46.6	2069766	66.5	1221367	40.4
Genotype	921628	26.2	101345	3.3	172070	5.7
Genotype × location	–	–	85633	2.8	61219	2.0
Genotype × year	266829	7.6	163231	5.2	156339	5.2
Genotype × location × year	–	–	278384	8.9	252138	8.3
Replication	72844	2.1	36548	1.2	458388	15.2
Incomplete block	56121	1.6	51841	1.7	40882	1.4
Residual	561969	16.0	225249	7.2	351171	11.6
Total	3520257	100.0	3112573	100.0	3021451	100.0

in 2014, while the environments with the least discriminative ability were YOU in 2010 and LE2 in 2014.

Genomic Predictions of Unphenotyped Genotypes

Predicting genotypic performance of new genotypes for each environment with the GBLUP_{G×E} was better than predicting for the mean of the environment with the GBLUP_M for most sets of environments (Fig. 4). The advantage of the GBLUP_{G×E} model was more evident for larger population sizes (Fig. 4, 5). Furthermore, LE1

(the location with the largest number of genotypes) had the largest difference in mean prediction comparing both GBLUP models (Fig. 5), GBLUP_{G×E} being better than the GBLUP_M. On the other hand, most of the sets grouping environments by years have higher accuracies with the GBLUP_M model (i.e., 2012, 2013, and 2014; Fig. 4, 5). In addition, 2012 shows the highest accuracy for the GBLUP_M model, with an accuracy of 0.5 (Fig. 5). Finally, no differences in prediction ability between GBLUP_M and GBLUP_{G×E} were found for large data sets ($N = 500$ or $N = 900$; Fig. 5).

Table 3. Additive main effect and multiplicative interaction (AMMI) and GGE family models for the genotype \times environment interaction analysis of wheat lines evaluated in 35 environments including locations in years. The sum of squares (SS), proportion of the total GEI variance, and test of significance for the first 10 principal components is shown.

Model	AMMI			GGE		
	SS	Proportion	P-value	SS	Proportion	P-value
1	1396.2	51.46	<0.0001	1889.8	44.02	<0.0001
2	616.6	22.72	<0.0001	1306.2	30.42	<0.0001
3	198.0	7.27	<0.0001	544.2	12.67	<0.0001
4	148.9	5.48	<0.0001	186.8	4.35	<0.0001
5	88.4	3.26	<0.0001	88.5	2.06	<0.0001
6	66.0	2.43	<0.0001	68.2	1.58	<0.0001
7	44.6	1.71	0.0003	47.3	1.10	0.0002
8	36.2	1.33	0.0332	36.3	0.84	0.0312
9	24.5	0.90	0.6731	27.1	0.63	0.4363
10	22.1	0.81	0.8108	23.3	0.54	0.7328

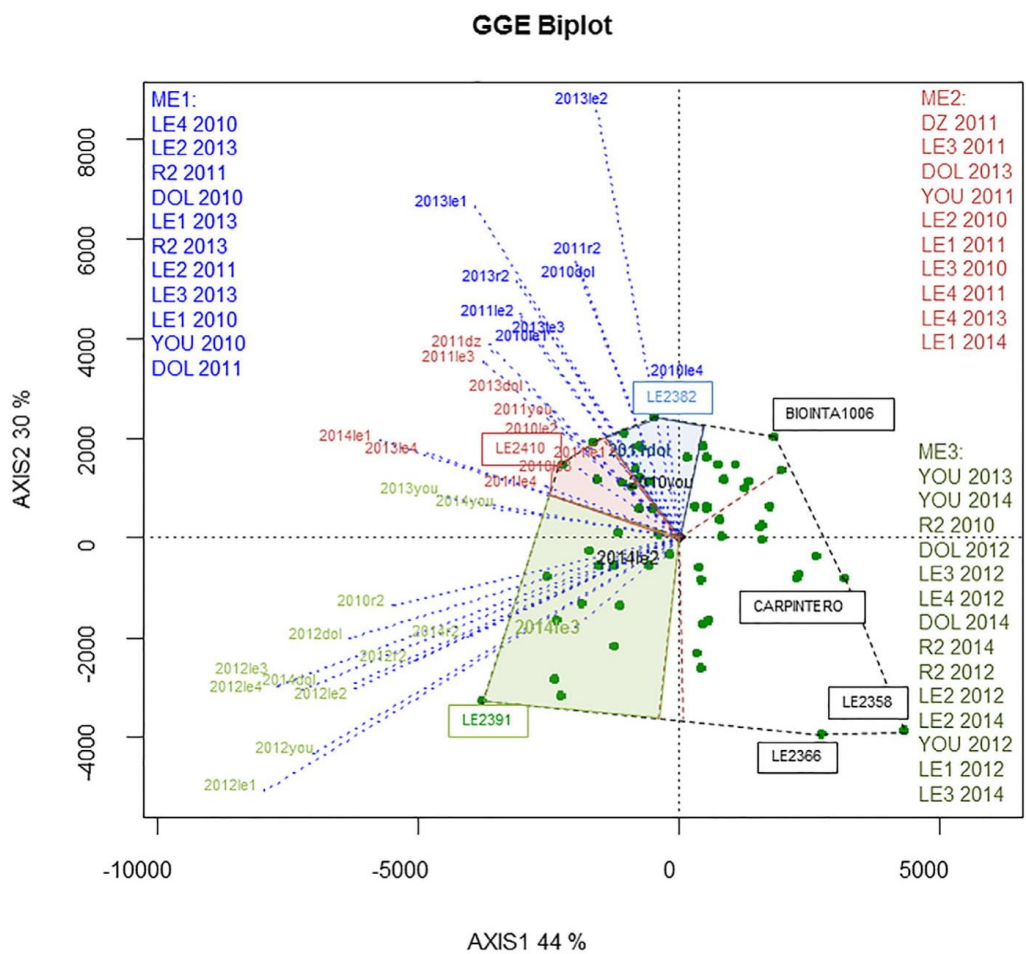


Figure 3. GGE biplot for yield in a wheat dataset considering 60 genotypes evaluated in 35 environments. Grouping in three mega-environments (MEs) based on winning genotypes (i.e., indicated within square) are shown (i.e., represented as polygon): ME1, ME2, and ME3. The environments for each ME are labeled outside the polygon.

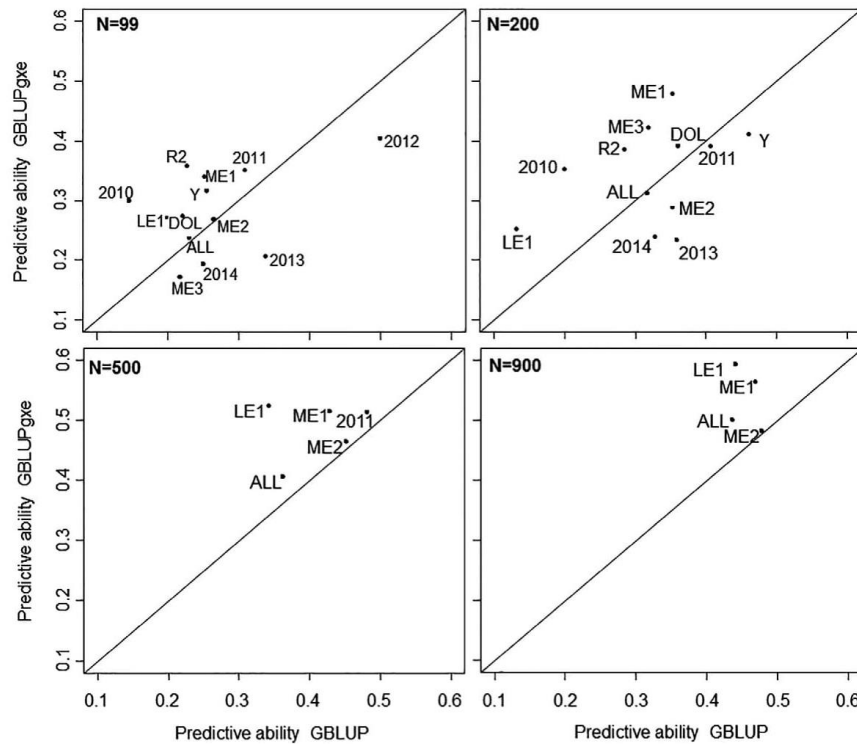


Figure 4. Accuracy of genomic breeding values predictions for yield in different sets of environments predicted either for the mean ($GBLUP_M$) or each environment within a set ($GBLUP_{g \times e}$) in 1477 wheat genotypes evaluated in 35 environments. The accuracy was estimated for different sample sizes ($N = 99$, $N = 200$, $N = 500$ and $N = 900$). Environments are grouped in sets by either years (i.e., 2010–2014), locations (i.e., DOL, LE1, R2, YOU), mega-environments (ME1–ME3), or all together (ALL).

Genomic Predictions of Unphenotyped Environments

When comparing independent training and testing sets and predicting for new environments, predictions for specific locations were high with either the $GBLUP_M$ or the $GBLUP_{g \times e}$ models using either information from other locations in a single year or using other environments within a ME (Table 4). However, predictions across years are poorly achieved using single locations with either $GBLUP$ models and are only reasonably achieved predicting within ME with either the $GBLUP_M$ or the $GBLUP_{g \times e}$ models (Table 4). The $GBLUP_M$ model should be used for predicting performance of genotypes in new environments (i.e., environments without previous phenotypic information) because the $GBLUP_{g \times e}$ model requires an estimation of genotypic correlations between environments that would be unavailable if no phenotypic information is available.

DISCUSSION

One of the most challenging issues in plant breeding is the use of highly unbalanced existing phenotyping in the presence of GEI (Burgueño et al., 2012; Dawson et al., 2013; Heslot et al., 2014). Breeding programs produce large MET datasets that are highly unbalanced and with different levels of quality because of the degree of replication in different trials (Bernardo, 2010). We used a large MET dataset with 1477 genotypes extensively phenotyped in 35 environments to compare different strategies for handling GEI. Our dataset had very good experimental design yield data quality with an average heritability on a line-mean basis of 0.82, and our predictive ability was moderate to high for a complex trait like grain yield with an overall prediction accuracy of 0.5. We specifically aimed to answer the following questions: Could we improve the predictive ability by modeling the variance-covariance matrix across environments and predicting performance of genotypes by environment? Could we design groups of environments were overall predictions

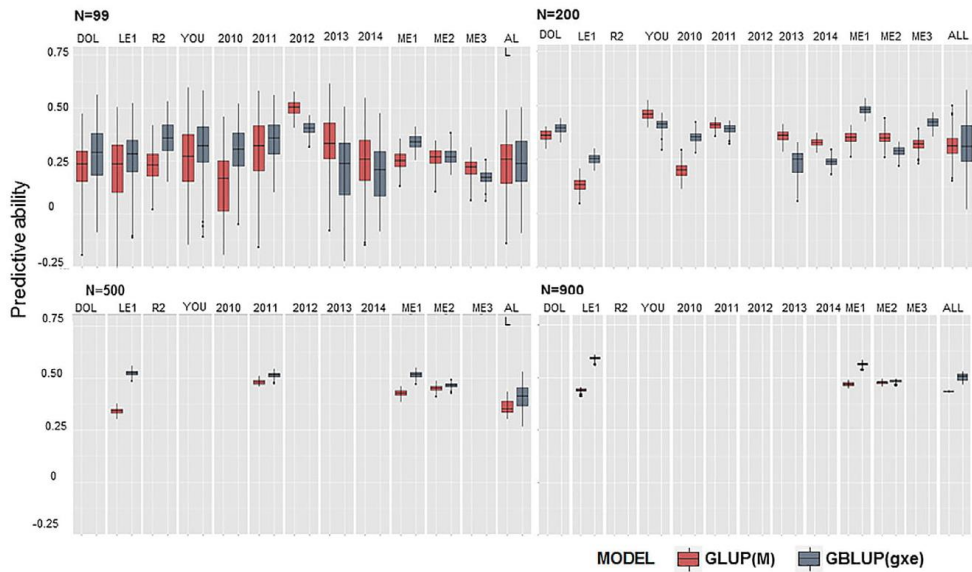


Figure 5. Accuracy of genomic breeding values predictions for yield in different sets of environments predicted either for the mean (GBLUP_M) or each environment within a set (GBLUP_{g×e}) in 1477 wheat genotypes evaluated in 35 environments indicating the standard deviation of the accuracy. The accuracy was estimated for different samples sizes ($N = 99$, $N = 200$, $N = 500$, and $N = 900$). Environments are grouped in sets by either years (i.e., 2010–2014), locations (i.e., Dolores [DOL], La Estanzuela [LE1], Young [YOU], Ruta2 [R2]), mega-environments (ME1–ME3) or all together (ALL). The range of prediction accuracy is for 100 replications of the sevenfold cross-validation.

Table 4. Accuracy for unphenotyped environments with GBLUP_M (1) and GBLUP_{g×e} (2), using independent training and testing sets for 1477 genotypes evaluated in 35 environments. Environments were grouped in sets by either year (2010–2014), location (Dolores [DOL], La Estanzuela [LE1], Young [YOU], Ruta2 [R2]) or mega-environment (ME1–ME4). Different training sets were used to predict genotypic performance in new environments (i.e. the testing set).

Set	Training set	Testing Year	Testing location							
			LE		YOU		DOL		R2	
			1	2	1	2	1	2	1	2
By year [†]	All locations in testing year but one	2010	0.31	0.31	0.12	0.17	0.71	0.71	0.57	0.44
		2011	0.42	0.39	0.39	0.31	0.52	0.37	0.44	0.57
		2012	0.74	0.79	0.44	0.54	0.78	0.81	0.57	0.65
		2013	0.03	0.05	0.40	0.74	0.36	0.59	0.28	0.36
		2014	0.23	-0.14	0.30	0.19	0.58	0.71	0.30	0.19
By location [‡]	All years for a location but one	2010	0.16	0.17	0.01	-0.02	-0.06	0.04	0.01	0.11
		2011	0.23	0.18	0.27	0.34	0.08	0.14	0.19	0.29
		2012	0.50	0.31	0.10	0.04	0.33	0.29	0.05	0.15
		2013	0.23	0.10	-0.05	0.08	0.31	0.56	0	0.19
		2014	0.21	0.20	0.06	-0.05	0.08	0.36	-0.05	0.33
By mega-environment [§]	All environments for a specific mega-environment but one	2010	0.10 (1) [¶]	0.20 (1)	0.11 (1)	0.18 (1)	0.72 (1)	0.72 (1)	0.32 (3)	0.35 (3)
		2011	0.36 (2)	0.35 (2)	0.36 (2)	0.44 (2)	0.39 (1)	0.28 (1)	0.38 (1)	0.39 (1)
		2012	0.63 (3)	0.45 (3)	0.46 (2)	0.54 (2)	0.75 (3)	0.80 (3)	0.63 (3)	0.30 (3)
		2013	0.28 (1)	0.18 (1)	0.46 (3)	0.43 (3)	0.43 (2)	0.50 (2)	0.44 (1)	0.32 (1)
		2014	0.24 (2)	0.25 (2)	-0.01 (2)	0.11 (2)	0.63 (3)	0.78 (3)	-0.03 (3)	0.20 (3)

[†] When sets of environments are grouped by year, the training set includes all genotypes in all locations but one for a given year, while the testing set includes the remaining location; one location at a time is left out of the training set and is being predicted.

[‡] When sets of environments are grouped by location, the training set includes all genotypes in all years but one for a given location, while the testing set includes the remaining year; 1 yr at a time is left out of the training set and is being predicted.

[§] When sets of environments are grouped by ME, the training set includes all genotypes in all environments within a ME but one, while the testing set includes the remaining environment; one environment at a time is left out of the training set and is being predicted.

[¶] Numbers 1 through 3 indicate the ME to which the specific environment belongs.

outperformed environment-specific predictions? Could we borrow information from other environments to predict performance in unphenotyped environments?

Correlation between Environments and Mega-Environments

Mixed Models are one of the most effective tools to handle unbalanced data while modeling correlation structures (Piepho, 1997, 1998; Yu et al., 2006; Piepho et al., 2007; Piepho and Möhring, 2007; Malosetti et al., 2007; Meyer, 2009). However, reliable estimates of correlation should be used to model the variance-covariance matrix among environments. In breeding programs, it is common to evaluate lines using less replication and environments during early testing (i.e., in PYT) saving resources for larger replication and environments of fewer lines in advanced trials (i.e., AYT or EYT; Bernardo, 2010). This creates a structure in GEI data where only promising lines are evaluated in multiple environments. Therefore, the estimation of the correlation among environments could be biased by the use of selected lines. We found that the genetic correlations matrices between environments estimated with 60 elite lines was similar to the correlation estimated with all available lines (i.e., $r = 0.7$, $p < 0.001$; Fig. 1; Supplemental Fig. S1). Furthermore, the genetic correlation between two environments for unselected lines (i.e., AYT) was similar to the correlation between the same two environments for selected lines (i.e., EYT; Supplemental Fig. S2). Piepho and Möhring (2006) argue that it is reasonable to assume that the change in allelic frequencies as a result of selection from training to testing populations is minimal, and therefore, prediction accuracy in the testing population is appropriate. Therefore, it is reasonable to assume that genotypic correlations among environments from selected individuals properly estimate correlations for the entire population. Other studies have shown that creating unbalanced data to sample more environments and genotypes produce higher prediction accuracy than having smaller population sizes and fewer environments with balanced data (Endelman et al., 2014).

Genotype \times Environment Interaction Characterization

Genotype \times environment interaction is a widespread phenomenon in plants, and therefore, a proper characterization of the GEI is key to decide on strategies to handle it (Crossa and Cornelius, 1997; Piepho, 1997, 1998; Burgueño et al., 2008; Malosetti et al., 2013). We found three MEs in our dataset: ME1 and ME2 have a strong correlation between environments and are composed of 2010, 2011, and 2013 environments; ME3 also has a strong correlation between environments and are composed of 2012 and 2014 environments. This pattern was detected by either the use of the best GGE or AMMI model, but

also by the GGE2 model using the winning genotypes in each ME (i.e., genotypes LE2382, LE2410, and LE2391; Fig. 3). Several studies have shown the advantages of either AMMI (Gauch, 2006, 2013; Gauch et al., 2008) or GGE models (Yan et al., 2000, 2007; Yang et al., 2009) to model GEI, but we identified the same general pattern with either an AMMI or a GGE model. We found that genotype \times year interactions contributed a large proportion of the GEI in our study. Year-to-year variations are more difficult to predict (Dawson et al., 2013) and usually contribute to a large proportion of the GEI (Ceretta and van Eeuwijk, 2008; Heslot et al., 2014). In our study, years 2010, 2011, and 2013 were high-yielding environments (i.e., yield average of 5370 kg ha⁻¹), while years 2012 and 2014 were low- or medium-yielding environments (i.e., yield average of 3946 kg ha⁻¹). Other studies found a similar pattern of environments grouping by yield performance (Byth et al., 1976; Windhausen et al., 2012). Furthermore, we found that while the average monthly rainfall in these locations is 80 mm mo⁻¹, heavy rainfall occurred during anthesis and grain filling in years 2012 and 2014 (Supplemental Fig. S3). The heavy rainfall increased disease presence, causing a severe infection of Fusarium head blight, which, in turn, decreased yield performance of lines. Further environmental characterization will provide a better insight into the GEI structure, but, by and large, we were able to successfully predict for rainy-dry anthesis period environments and might be able to predict performance for Fusarium and non-Fusarium years in this dataset.

Predictive Ability

Plant breeders seek to develop varieties with the best performance for their target environments (Comstock, 1977). Most research on genomic selection has focused on accurately predicting the genotypic mean performance over all environments, assuming that all trial locations belong to the same target population of environments (Asoro et al., 2011; Charmet and Storlie, 2012; Storlie and Charmet, 2013; Crossa et al., 2013; Dawson et al., 2013). However, since GEI is widespread, the accuracy of genomic prediction could be lowered especially by the presence of crossover interaction (Burgueño et al., 2012; Dawson et al., 2013; Jarquín et al., 2014). Mixed models are powerful tools to make predictions by modeling the variance-covariance structure across environments to account for GEI (Piepho, 1997; Burgueño et al., 2008; Meyer, 2009; Malosetti et al., 2013). We compared two strategies to deal with GEI: reduce it (i.e., predict overall performance for a set of environments, GBLUP_M) and exploit it (i.e., predict performance by environment using the correlations among environments, GBLUP_{g \times e}). We obtained high overall predictions accuracies for grain yield (GBLUP_M) compared with other studies (Heffner et al., 2011; Burgueño et al., 2012; Poland et al., 2012b; Charmet and

Storlie, 2012; Dawson et al., 2013; Lado et al., 2013; Crossa et al., 2014; Lopez-Cruz et al., 2015). This probably is due to the large population size with extensive phenotyping and high quality experimental design with two to four replications and high heritability. High prediction ability has been found to be associated with large population sizes, a large number of sampled environments (Riedelsheimer and Melchinger, 2013) even at the cost of unbalanced designs (Heffner et al., 2011; Endelman et al., 2014), and good experimental designs (Cooper et al., 2014) with large numbers of replications (Lorenz, 2013).

To reduce GEI, environments were grouped into sets based on different criteria (i.e., year, location, or ME). The genotype \times environment model (i.e., GBLUP_{g \times e}) outperformed the overall model (GBLUP_M) when predicting within MEs or locations. On the other hand, most of the overall predictions within years (i.e., GBLUP_M) were higher than by-environment predictions (i.e., GBLUP_{g \times e}), indicating low GEI within years. The only exception to this pattern is the location YOU. The improvement in the predictive ability of the GBLUP_{g \times e} can be explained by the shared information among environments. The GBLUP_M model is equivalent to the GBLUP_{g \times e} model with a correlation between environments equal to one. Therefore, when correlations among environments are high (like correlations among locations within years), similar predictive ability is found between GBLUP_M and GBLUP_{g \times e} models.

Furthermore, we found that predicting unphenotyped environments is better achieved within MEs and years. This result is not unexpected because lower GEI is found within years or MEs, while higher GEI is found across years. The use of multienvironment models to make overall genomic predictions has been documented elsewhere (Jarquín et al., 2014). However, the use of predictions within MEs has improved predictive ability in some studies (Burgueño et al., 2012; Dawson et al., 2013) but not in others (Windhausen et al., 2012). We found that grouping environments in MEs that are mainly based on year-to-year repeatable variations and predicting overall performance yielded high predictive ability in most of the situations. Therefore, using locations to sample GEI is not a good strategy to capture all GEI for this program. Furthermore, predictions based on single-year phenotypic evaluation were unreliable, and historic data from breeding programs in the form of MET data is the key to characterize GEI and to increase predictive ability of GS models. However, we found that even random environmental factors (i.e., years) could be better modeled when GEI is included in the prediction model with the GBLUP_{g \times e} model. Furthermore, when the size of population increases, all sets of environments and models tended to have the same performance with close predictive ability.

We were able to successfully predict new genotypes borrowing information from relatives evaluated in

multiple environments using the GBLUP_{g \times e} model. This model proved to be superior for predicting unphenotyped genotypes. However, where genomic predictions for new environments is desirable, the use of genetic correlations between environments is not possible, and therefore the GBLUP_M should be used. We showed that for predicting performance of genotypes in new environments, the best strategy was to either predict across locations for single years or to predict within ME for any year or location.

CONCLUSIONS

Modeling GEI increased predictive ability. The use of MEs to identify groups of environments with small GEI, and overall predictions within MEs produces higher prediction accuracy compared with overall predictions across years. Borrowing information from environments by modeling the variance-covariance matrix is beneficial for predicting new genotypes before phenotyping. Finally, when large data sets were used, the sets of environments and models tend to perform similar.

Acknowledgments

We would like to thank three anonymous reviewers for their comments that contributed to this manuscript. We would also like to thank Jesse Poland for his review of this manuscript. Funding for this project was provided by INIA (Project L2-2011) and the Research and Innovation Agency of Uruguay (ANII), through Grant POS_NAC_2013_1_11261. We express our appreciation for the effort of the technical personnel of INIA wheat breeding program Leonardo Hernández, María Ferreira, Dumas Laun, Abel López, and José Flores.

References

- Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, and J.L. Janink. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *Plant Gen.* 4:132–144. doi:10.3835/plantgenome2011.02.0007
- Bennewitz, J., T. Solberg, and T. Meuwissen. 2009. Genomic breeding value estimation using nonparametric additive regression models. *Genet. Sel. Evol.* 41:20–32. doi:10.1186/1297-9686-41-20
- Bernardo, R. 2010. *Breeding for quantitative traits in plants*. 2nd ed. Stemma Press, Woodbury, MN. ISBN 978-0-9720724-1-0.
- Boer, M.P., D. Wright, L. Feng, D.W. Podlich, L. Luo, M. Cooper, and F.A. van Eeuwijk. 2007. A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. *Genetics* 177:1801–1813. doi:10.1534/genetics.107.071068
- Braun, H.J., S. Rajaram, and M. van Ginkel. 1996. CIMMYT's approach to breeding for wide adaptation. *Euphytica* 92:175–183. doi:10.1007/BF00022843
- Burgueño, J., J. Crossa, P.L. Cornelius, and R.C. Yang. 2008. Using factor analytic models for joining environments and genotypes without crossover genotype \times environment interaction. *Crop Sci.* 48:1–15. doi:10.2135/cropsci2007.11.0632

- Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa. 2012. Genomic prediction of breeding values when modeling genotype \times environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52:707–719. doi:10.2135/cropsci2011.06.0299
- Byth, D., R. Eisemann, and I. DeLacy. 1976. Two-way pattern analysis of a large data set to evaluate genotypic adaptation. *Heredity (Edinb)* 37:215–230. doi:10.1038/hdy.1976.84
- Ceretta, S., and F.A. van Eeuwijk. 2008. Grain yield variation in malting barley cultivars in Uruguay and its consequences for the design of a trials network. *Crop Sci.* 48:167–180. doi:10.2135/cropsci2006.06.0428
- Charmet, G., and E. Storlie. 2012. Implementation of genome-wide selection in wheat. *Russ. J. Genet. Appl. Res.* 2:298–303. doi:10.1134/S207905971204003X
- Comstock, R.E. 1977. Quantitative genetics and the design of breeding programs. In: E. Pollack, O. Kempthorne, T.B. Bailey Jr., editors, *Proceedings of the International Conference on Quantitative Genetics*. 16–21 Aug. 1976. Iowa State Univ. Press, Ames, IA. p. 705–718.
- Cooper, M., and I.H. Delacy. 1994. Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theor. Appl. Genet.* 88:561–572. doi:10.1007/BF01240919
- Cooper, M., C. Messina, and D. Podlich. 2014. Predicting the future of plant breeding: Complementing empirical evaluation with genetic prediction. *Crop Pasture Sci.* 65:311–336. doi:10.1071/CP14007
- Crossa, J. 2012. From genotype \times environment interaction to gene \times environment interaction. *Curr. Genomics* 13:225–244. doi:10.2174/138920212800543066
- Crossa, J., Y. Beyene, S. Kassa, P. Pérez, J.M. Hickey, C. Chen, G. de los Campos, J. Burgueño, V.S. Windhausen, E. Buckler, J.L. Jannink, M. a Lopez Cruz, and R. Babu. 2013. Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3: Genes, Genomes, Genet.* 3:1903–1926.
- Crossa, J., and P.L. Cornelius. 1997. Sites regression and shifted multiplicative model clustering of cultivar trial sites under heterogeneity of error variances. *Crop Sci.* 37:406–415. doi:10.2135/cropsci1997.0011183X003700020017x
- Crossa, J., P. Pérez, J. Hickey, and J. Burgueño. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity (Edinb)* 112:48–60. doi:10.1038/hdy.2013.16
- Cullis, B.R., A.B. Smith, C.P. Beeck, and W.A. Cowling. 2010. Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. *Genome* 53:1002–1016. doi:10.1139/G10-080
- Dawson, J.C., J.B. Endelman, N. Heslot, J. Crossa, J. Poland, S. Dreisigacker, Y. Manès, M.E. Sorrells, and J.L. Jannink. 2013. The use of unbalanced historical data for genomic selection in an international wheat breeding program. *F. Crop. Res.* 154:12–22. doi:10.1016/j.fcr.2013.07.020
- DeLacy, I.H., K.E. Basford, M. Cooper, J.K. Bull, and C.G. McLaren. 1996. Analysis of multi environment trials: An historical perspective. *Plant adaptation and crop improvement*. 39–124.
- de los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193:327–345. doi:10.1534/genetics.112.143313
- Dreccer, M.F., M.G. Borgognone, F.C. Ogbonnaya, R.M. Trethowan, and B. Winter. 2007. CIMMYT-selected derived synthetic bread wheats for rainfed environments: Yield evaluation in Mexico and Australia. *F. Crop. Res.* 100:218–228. doi:10.1016/j.fcr.2006.07.005
- Eisemann, R.L., M. Cooper, and D.R. Woodruff. 1990. Beyond the analytical methodology: Better interpretation of genotype by environment interaction. In: M.S. Kang, editor, *Genotype by environment interaction and plant breeding*. Louisiana State University, Baton Rouge, LA. p. 108–117.
- Elshire, R., J. Glaubitz, and Q. Sun. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi:10.1371/journal.pone.0019379
- Endelman, J. 2012. Using rrBLUP 4.0. <http://www.afhalifax.ca/bete/revive-eponge/Rblkfct/vignette.pdf> (accessed 28 Oct. 2015).
- Endelman, J.B., G.N. Atlin, Y. Beyene, K. Semagn, X. Zhang, M.E. Sorrells, and J.L. Jannink. 2014. Optimal design of preliminary yield trials with genome-wide markers. *Crop Sci.* 54:48–59. doi:10.2135/cropsci2013.03.0154
- Frutos, E., M.P. Galindo, and V. Leiva. 2014. An interactive biplot implementation in R for modeling genotype-by-environment interaction. *Stoch. Environ. Res. Risk Assess.* 28:1629–1641.
- Gabriel, K.R. 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58:453–467. doi:10.1093/biomet/58.3.453
- Gauch, H.G. 2006. Statistical analysis of yield trials by AMMI and GGE. *Crop Sci.* 46:1488–1500. doi:10.2135/cropsci2005.07-0193
- Gauch, H.G. 2013. A simple protocol for AMMI analysis of yield trials. *Crop Sci.* 53:1860–1869. doi:10.2135/cropsci2013.04.0241
- Gauch, H.G., H.P. Piepho, and P. Annicchiarico. 2008. Statistical analysis of yield trials by AMMI and GGE: Further considerations. *Crop Sci.* 48:866–889. doi:10.2135/cropsci2007.09.0513
- Gauch, H.G., and R.W. Zobel. 1988. Predictive and postdictive success of statistical analyses of yield trials. *Theor. Appl. Genet.* 76:1–10. doi:10.1007/BF00288824
- Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, and E.S. Buckler. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9:E90346. doi:10.1371/journal.pone.0090346
- Gollob, H. 1968. A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika* 33:73–115. doi:10.1007/BF02289676
- Gutiérrez, L., S. Germán, S. Pereyra, P.M. Hayes, C.a. Pérez, F. Capettini, A. Locatelli, N.M. Berberian, E.E. Falconi, R. Estrada, D. Fros, V. Gonza, H. Altamirano, J. Huerta-Espino, E. Neyra, G. Orjeda, S. Sandoval-Islas, R. Singh, K. Turkington, and A.J. Castro. 2015. Multi-environment multi-QTL association mapping identifies disease resistance QTL in barley germplasm from Latin America. *Theor. Appl. Genet.* 128:501–516. doi:10.1007/s00122-014-2448-y
- Hayes, B.J., N.O.I. Cogan, L.W. Pembleton, M.E. Goddard, J. Wang, G.C. Spangenberg, and J.W. Forster. 2013. Prospects for genomic selection in forage plant species. *Plant Breed.* 132:133–143. doi:10.1111/pbr.12037
- Heffner, E.L., J. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Gen.* 4:65–75. doi:10.3835/plantgenome.2010.12.0029

- Heffner, E.L., A.J. Lorenz, J.L. Jannink, and M.E. Sorrells. 2010. Plant breeding with genomic selection: Gain per unit time and cost. *Crop Sci.* 50:1681–1690. doi:10.2135/cropsci2009.11.0662
- Heffner, E.L., M.E. Sorrells, and J.L. Jannink. 2009. Genomic selection for crop improvement. *Crop Sci.* 49:1–12. doi:10.2135/cropsci2008.08.0512
- Henderson, C.R. 1984. Applications of linear models in animal breeding models. 3rd ed. University of Guelph, Guelph, ON.
- Heslot, N., D. Akdemir, M.E. Sorrells, and J.L. Jannink. 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127:463–480. doi:10.1007/s00122-013-2231-5
- Heslot, N., H.P. Yang, M.E. Sorrells, and J.L. Jannink. 2012. Genomic selection in plant breeding: A comparison of models. *Crop Sci.* 52:146–160. doi:10.2135/cropsci2011.06.0297
- Isidro, J., J.L. Jannink, D. Akdemir, J. Poland, N. Heslot, and M.E. Sorrells. 2014. Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* 128:145–158.
- Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt, J. Lorgeou, F. Piroux, L. Guerreiro, P. Pérez, M. Calus, J. Burgueño, and G. de los Campos. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127:595–607. doi:10.1007/s00122-013-2243-1
- Kelly, A.M., B.R. Cullis, A.R. Gilmour, J.A. Eccleston, and R. Thompson. 2009. Estimation in a multiplicative mixed model involving a genetic relationship matrix. *Genet. Sel. Evol.* 41:33–42. doi:10.1186/1297-9686-41-33
- Lado, B., I. Matus, A. Rodríguez, L. Inostroza, J. Poland, F. Belzile, A. del Pozo, M. Quincke, M. Castro, and J. von Zitzewitz. 2013. Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3: Genes, Genomes, Genet.* 3:2105–2114.
- Lal, R.K. 2013. Adaptability patterns and stable cultivar selection in menthol mint (*Mentha arvensis* L.). *Ind. Crops Prod.* 50:176–181. doi:10.1016/j.indcrop.2013.07.008
- Lal, R.K. 2014. Breeding for new chemotypes with stable high essential oil yield in *Ocimum*. *Ind. Crops Prod.* 59:41–49. doi:10.1016/j.indcrop.2014.04.047
- Legarra, A., C. Robert-Granié, E. Manfredi, and J.M. Elsen. 2008. Performance of genomic selection in mice. *Genetics* 180:611–618. doi:10.1534/genetics.108.088575
- Lopez-Cruz, M., J. Crossa, D. Bonnett, S. Dreisigacker, J. Poland, J.L. Jannink, R.P. Singh, E. Auriague, and G. de los Campos. 2015. Increased prediction accuracy in wheat breeding trials using a marker \times environment interaction genomic selection model. *G3: Genes, Genomes, Genet.* 5:569–582.
- 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., K. Smith, and J. Jannink. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. *Crop Sci.* 52:1609–1621. doi:10.2135/cropsci2011.09.0503
- Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor. Appl. Genet.* 120:151–161. doi:10.1007/s00122-009-1166-3
- Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch, R. Okechukwu, A.G.O. Dixon, P. Kulakow, and J.L. Jannink. 2013. Relatedness and genotype \times environment interaction affect prediction accuracies in genomic selection: A study in cassava. *Crop Sci.* 53:1312–1325. doi:10.2135/cropsci2012.11.0653
- Malosetti, M., J.M. Ribaut, and F.A. van Eeuwijk. 2013. The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. *Front. Physiol.* 4:44.
- Malosetti, M., J.M. Ribaut, M. Vargas, J. Crossa, and F.A. van Eeuwijk. 2007. A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L.). *Euphytica* 161:241–257. doi:10.1007/s10681-007-9594-0
- Malosetti, M., J. Voltas, I. Romagosa, S.E. Ullrich, and F.A. van Eeuwijk. 2004. Mixed models including environmental covariables for studying QTL by environment interaction. *Euphytica* 137:139–145. doi:10.1023/B:EUPH.0000040511.46388.ef
- Mandel, J. 1969. The partitioning of interaction in analysis of variance. *J. Res. Natl. Bur. Stand., Sect. B* 73B:309–328.
- Mathews, K.L., M. Malosetti, S. Chapman, L. McIntyre, M. Reynolds, R. Shorter, and F.A. van Eeuwijk. 2008. Multi-environment QTL mixed models for drought stress adaptation in wheat. *Theor. Appl. Genet.* 117:1077–1091. doi:10.1007/s00122-008-0846-8
- Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829.
- Meyer, K. 2009. Factor-analytic models for genotype \times environment type problems and structured covariance matrices. *Genet. Sel. Evol.* 41:21–32. doi:10.1186/1297-9686-41-21
- Patterson, H., and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika* 58:545–554. doi:10.1093/biomet/58.3.545
- Piepho, H.P. 1997. Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics* 53:761–766. doi:10.2307/2533976
- Piepho, H.P. 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theor. Appl. Genet.* 97:195–201. doi:10.1007/s001220050885
- Piepho, H.P. 2000. A mixed-model approach to mapping quantitative trait loci in Barley on the basis of multiple environment data. *Genetics* 156:2043–2050.
- Piepho, H.P. 2005. Statistical tests for QTL and QTL-by-environment effects in segregating populations derived from line crosses. *Theor. Appl. Genet.* 110:561–566. doi:10.1007/s00122-004-1872-9
- Piepho, H.P., and J. Möhring. 2006. Selection in cultivar trials: Is it ignorable? *Crop Sci.* 46:192–201. doi:10.2135/cropsci2005.04-0038
- Piepho, H.P., and J. Möhring. 2007. Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* 177:1881–1888. doi:10.1534/genetics.107.074229
- Piepho, H.P., J. Möhring, A.E. Melchinger, and A. Büchse. 2007. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161:209–228. doi:10.1007/s10681-007-9449-8
- Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team. 2013. nlme: Linear and nonlinear mixed effects models. R package version 3.1-122. <http://CRAN.R-project.org/package=nlme> (accessed 1 Feb. 2015).
- Poland, J.A., P.J. Brown, M.E. Sorrells, and J.L. Jannink. 2012a. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* 7:E32253. doi:10.1371/journal.pone.0032253

- Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.L. Jannink. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Gen.* 5:103–113. doi:10.3835/plantgenome2012.06.0006
- R Development Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rebonato, R., and P. Jäckel. 1999. The most general methodology to create a valid correlation matrix for risk management and option pricing purposes. *Appl. Econ. Lett.* 17:1767–1768.
- Resende, M.F.R., P. Muñoz, J.J. Acosta, G.F. Peter, J.M. Davis, D. Grattapaglia, M.D.V. Resende, and M. Kirst. 2011. Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. *New Phytol.* 193:617–624. doi:10.1111/j.1469-8137.2011.03895.x
- Riedelsheimer, C., and A.E. Melchinger. 2013. Optimizing the allocation of resources for genomic selection in one breeding cycle. *Theor. Appl. Genet.* 126:2835–2848. doi:10.1007/s00122-013-2175-9
- Rodrigues, P.C., M. Malosetti, H.G. Gauch, and F.A. van Eeuwijk. 2014. A weighted AMMI algorithm to study genotype-by-environment interaction and QTL-by-environment interaction. *Crop Sci.* 54:1555–1570. doi:10.2135/cropsci2013.07.0462
- Rutkoski, J., R.P. Singh, J. Huerta-Espino, S. Bhavani, J. Poland, J.L. Jannink, and M.E. Sorrells. 2015. Efficient use of historical data for genomic selection: A case study of stem rust resistance in wheat. *Plant Gen.* 8:1–10. doi:10.3835/plantgenome2014.09.0046
- SAS Institute. 2005. JMP user guide. Release 6. SAS Institute Inc., Cary, NC.
- Solberg, T.R., A.K. Sonesson, J.A. Woolliams, and T.H.E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. *Genet. Sel. Evol.* 41:29–37. doi:10.1186/1297-9686-41-29
- Storlie, E., and G. Charmet. 2013. Genomic selection accuracy using historical data generated in a wheat breeding program. *Plant Gen.* 6:1–9. doi:10.3835/plantgenome2013.01.0001
- van Eeuwijk, F.A. 1995. Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. *Euphytica* 84:1–7. doi:10.1007/BF01677551
- van Eeuwijk, F.A., M. Malosetti, X. Yin, P.C. Struik, and P. Stam. 2005. Statistical models for genotype by environment data: From conventional ANOVA models to eco-physiological QTL models. *Aust. J. Agric. Res.* 56:883–894. doi:10.1071/AR05153
- Wang, H., and P. Schmidt. 2002. One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *J. Prod. Anal.* 18:129–144. doi:10.1023/A:1016565719882
- Windhausen, V.S., S. Wagener, C. Magorokosho, D. Makumbi, B. Vivek, H.P. Piepho, A.E. Melchinger, and G.N. Atlin. 2012. Strategies to subdivide a target population of environments: Results from the CIMMYT-Led maize hybrid testing programs in Africa. *Crop Sci.* 52:2143–2152. doi:10.2135/cropsci2012.02.0125
- Yan, W., L. Hunt, Q. Sheng, and Z. Szlavncics. 2000. Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Sci.* 40:597–605. doi:10.2135/cropsci2000.403597x
- Yan, W., M.S. Kang, B. Ma, S. Woods, and P.L. Cornelius. 2007. GGE biplot vs. AMMI analysis of genotype-by-environment data. *Crop Sci.* 47:643–655. doi:10.2135/cropsci2006.06.0374
- Yan, W., and N. Tinker. 2006. Biplot analysis of multi-environment trial data: Principles and applications. *Can. J. Plant Sci.* 6:623–645. doi:10.4141/P05-169
- Yang, R.C., J. Crossa, P.L. Cornelius, and J. Burgueño. 2009. Biplot analysis of genotype \times environment interaction: Proceed with caution. *Crop Sci.* 49:1564–1576. doi:10.2135/cropsci2008.11.0665
- Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38:203–208. doi:10.1038/ng1702
- Zobel, R.W., M.J. Wright, and H.G. Gauch, Jr. 1988. Statistical analysis of a yield trial. *Agron. J.* 80:388–393. doi:10.2134/agronj1988.00021962008000030002x

4. OPTIMIZACIÓN DE RECURSOS EN LA PREDICCIÓN DE MÚLTIPLES CARACTERES DE CALIDAD PANADERA DE TRIGO

4.1. RESUMEN DEL CAPÍTULO

Las líneas experimentales dentro de un programa de mejoramiento son evaluadas para varios caracteres por los cuales se buscan mejorar. Muchos de estos caracteres están correlacionados entre sí, dado que en su expresión intervienen los mismos genes, tal es el caso de varios de los caracteres de calidad panadera en trigo. La predicción conjunta de estos caracteres, mediante modelos multi-carácter, puede utilizarse para fenotipar menos caracteres de alto costo y difícil medición y predecirlos a través de caracteres correlacionados de menor costo y/o de determinación más sencilla. El uso de modelos multi-carácter podría ser una ventaja en la optimización de recursos dentro de un programa de mejoramiento. En este trabajo se utilizaron modelos multi-carácter para evaluar qué proporción del fenotipado de un carácter es posible reducir si se compensa utilizando caracteres correlacionados fenotipados en toda la población de entrenamiento y predicción. Luego, esto mismo se evaluó utilizando el 50% de los individuos fenotipados, para los caracteres correlacionados, de forma balanceada y desbalanceada. Se fenotiparon y genotiparon 495 individuos para ocho caracteres de calidad panadera y 6.655 SNP utilizando genotipado por secuenciación. Se obtuvieron dos grupos de cuatro caracteres más correlacionados entre sí que fueron predichos por separado. Utilizando los modelos multi-carácter fue posible reducir hasta un 30% el fenotipado de un carácter si se cuenta con el 100% de las líneas fenotipadas para caracteres correlacionados. Fenotipar el 50% de las líneas para caracteres correlacionados de forma desbalanceada es una buena estrategia que permite optimizar el uso de recursos e incrementar las predicciones respecto al modelo de un solo carácter. Concluimos que los modelos multi-carácter son una herramienta útil en la optimización de recursos de un programa de mejoramiento, dado que es posible reducir significativamente el fenotipado de un carácter de alto costo y difícil medición, si se cuenta con caracteres correlacionados.

**RESOURCE ALLOCATION OPTIMIZATION WITH MULTI-TRAIT
GENOMIC PREDICTION FOR BREAD WHEAT (*Triticum aestivum* L.)
BAKING QUALITY**

Bettina Lado, Daniel Vázquez, Martin Quincke, Paula Silva, Ignacio Aguilar and
Lucia Gutiérrez*

B. Lado, and L. Gutiérrez, Department of Statistics, Facultad de Agronomía, Universidad de la República, Garzón 780, Montevideo 12900, Uruguay; L. Gutiérrez, Department of Agronomy, University of Wisconsin– Madison, 1575 Linden Dr, Madison, WI 53706, USA; D.Vázquez, M. Quincke, P. Silva Instituto Nacional de Investigación Agropecuaria, Est. Exp. La Estanzuela, Ruta 50 km 11.5, Colonia, 70006, Uruguay; I. Aguilar, Instituto Nacional de Investigación Agropecuaria, Est. Exp. Las Brujas, Ruta 48 km 10, Rincón del Colorado, Canelones 90200, Uruguay

*Corresponding authors: gutierrezcha@wisc.edu.

ABSTRACT

Multi-trait genomic prediction models can be used to predict labor intensive or expensive correlated traits where phenotyping depth of correlated traits could be larger than phenotyping depth of targeted traits, reducing resources and improving prediction accuracy. This is particularly important in the context of allocating phenotyping resources in plant breeding programs. The objective of this work was to evaluate multi-trait models' predictive ability with different depth of phenotypic information from correlated traits. We evaluated 495 wheat advanced breeding lines for eight baking quality traits which were genotyped with genotyping-by-sequencing. Through different approaches for cross-validation we evaluated the predictive ability of a single-trait model and a multi-trait model. Moreover, we evaluated different sizes of the training population (from 50 to 396 individuals) for the trait of interest, different depth of phenotypic information for correlated traits (50 and 100%), and the number of correlated traits to be used (one to three). There was no loss in the predictive ability by reducing the training population up to 30% (149 individuals) when using correlated traits. The use of a multi-trait model with one highly correlated trait phenotyped for both, the training and the testing sets, was the best model considering phenotyping resources and gain in predictive ability. The inclusion of correlated traits in the training population and testing lines is a strategic approach to replace phenotyping of labor intensive and high cost traits in a breeding program.

Keywords: genomic selection, unbalanced phenotyping, correlated traits.

AUTHOR CONTRIBUTION STATEMENT

Designed the phenotyping experiments: DV and MQ. Performed genotyping analyses: BL and PS. Performed statistical analyses: BL, LG and IA. Wrote the paper: BL, LG, PS and IA. Designed the study and hypothesis: LG. All authors read and approved the final manuscript.

Key Message: Multi-trait genomic prediction models are useful to allocate available resources in breeding programs by targeted phenotyping of correlated traits when predicting expensive and labor intensive quality parameters.

ACKNOWLEDGEMENTS

We express our appreciation for the effort of the technical personnel of INIA La Estanzuela from “Laboratorio de calidad industrial de granos”. Support for doctoral work of BL was provided by Agencia Nacional de Investigación e Innovación (ANII) - Uruguay, through Grant POS_NAC_2013_1_11261 and by Comisión Sectorial de Investigación Científica (CSIC) - Uruguay, through grants in the program internships abroad. We would like to thank two anonymous reviewers for their comments that improved the manuscript.

4.2. INTRODUCTION

Wheat is one of the most important staple food crop of humans (Shewry and Hey 2015), providing 18% of the total caloric intake of the world (FAO 2017). Bread is one of the most important end use products of wheat, therefore improving bread quality is a key aspect of wheat improvement. Baking bread quality is a complex trait with quantitative inheritance, derived by several individual traits, each one with a different level of environmental influence (Williams et al. 2008). One of the traits that determines baking quality is gluten strength, which is a key factor for loaf volume that determines rising and shape maintenance during the baking process (MacRitchie 1992). Gluten strength is affected by the proportion of glutenins and gliadins polypeptides (75-80% of total proteins) synthesized by each wheat variety (MacRitchie 1992). Therefore, the amount and quality of proteins determines gluten strength and can be evaluated with the sedimentation volume value (MacRitchie 1992). Protein quantity also determines the stability of the dough to create a network and retain water, which is called wet gluten (WG). There is a complex interaction between proteins and other components, like pentosans (Hamer et al. 2009), making dough strength predictability from chemical composition very difficult. Therefore, rheological tests are required, including the use of the alveograph to investigate the stretching properties of the dough. The total energy required for breaking a standardized bubble is called baking strength (W). The length of the curve (L), or the time required to break it, is the extensibility and the height of the peak (P) represents the tenacity or maximum dough resistance to rupture (Indrani et al. 2007). All these traits are used to select varieties with better bread quality (Vázquez 2009).

Baking quality is a complex quantitative trait and several breeding strategies have been successfully used to improve complex traits. In the pre-genomic era common breeding strategies involved the use of classical quantitative genetic approaches (Lynch and Walsh 1998), including pedigree information, to estimate Best Linear Unbiased Predictors (BLUPs; Henderson and Quaas 1976). With the widespread availability of molecular markers, genomic approaches have been used (Lande and Thompson 1990). One of the most widely used strategies involves using the additive relationship matrix estimated from markers instead of the additive

relationship matrix estimated from pedigree with BLUP models. This was the beginning of the genomic selection (GS) era, and the new BLUP model was called a G-BLUP model (Meuwissen et al. 2001; Habier et al. 2007). The G-BLUP model is equivalent to a Ridge Regression BLUP model (RR-BLUP; Habier et al. 2007) and to Bayesian models that assume Gaussian priors for marker effects (i.e. Bayesian Ridge Regression; VanRaden 2008). The latter models compute genetic breeding values (GEBV) adding up all marker effects which were estimated assuming a normal distribution of marker effects with common variance. These models have been the model of choice in most of the prediction scenarios due to their high predictive ability and simple implementation (Heslot et al. 2015). However, other models that assume heterogeneity of variances and different distributions for marker effects performing variable selection (de los Campos et al. 2013) have higher predictive ability in specific situations (reviewed in Lorenz et al. 2011).

Classical quantitative genetics theory also provides the necessary framework for selecting multiple traits; classical approaches used for multi-trait selection include an independent culling approach, tandem and index selection (Falconer and Mackay 1996). Multi-trait (MT) selection is justified only when traits are genetically correlated (Henderson and Quaas 1976) which can be the result of pleiotropy or linkage disequilibrium between genes (Falconer and Mackay 1996). A particular case of selection indices (i.e. Smith-Hazel Index (Smith 1936; Hazel 1943)) use the correlation and the phenotypic and genotypic variance-covariance matrices among traits to estimate the net merit of genotypes. Multi-trait predictions have been extended for the use of genomic information (Calus and Veerkamp 2011; Ceron-Rojas et al. 2015). In this case, the correlation among observations is a function of the additive genetic correlation among traits and the additive genetic relationship among individuals (Calus and Veerkamp 2011). This same strategy was applied to predict lines in the context of genotype by environment interaction using the environmental correlation as multi-trait (Burgueño et al. 2012). Similar to single-trait models, multi-trait models could assume different marker distributions to estimate the breeding values. Calus and Veerkamp (2011) presented three models, one model estimates all marker effects assuming a unique variance (multi-trait G-BLUP); the other two models perform variable

selection, one assumes the same variance for all the SNPs (BayesC π), while the other assumes different variances whether the SNPs were or not associated with a Quantitative Trait Loci (QTL; BayesSSVS). In this study the differences in performance among methods was small (Calus and Veerkamp 2011). Jia and Jannink (2012) compared three models that were similar to those of Calus and Veerkamp (2011) but used different genetic architectures for the traits, and found that the models assuming different variance (BayesA) and performing variable selection (BayesC π) were the best option when QTL of major effects were simulated. However, for truly quantitative traits, models assuming normal distribution of marker effects and unique variances were similar to more complex models. Therefore, a simple model using additive relationship matrix as variance–covariance matrix among individuals seems to perform well for predicting single (Habier et al. 2007) and multi-trait models (Jia and Jannink 2012; Guo et al. 2014).

Prediction of new un-phenotyped individuals has been studied using multi-trait and single-trait genomic prediction models (Calus and Veerkamp 2011; Jia and Jannink 2012; Guo et al. 2014). Calus and Veerkamp (2011) did not find significant differences between both models using simulated data for predicting traits with different but high heritability ($h^2=0.6$ and $h^2=0.9$). However, the multi-trait model showed good performance to predict low heritability traits with the help of correlated traits with high heritability (Jia and Jannink 2012; Guo et al. 2014), and was optimal when the genetic architecture was explained by major QTL. Finally, the advantage of multi-trait models to predict new un-phenotyped individuals was not so obvious when using experimental data for disease resistance in pine (Jia and Jannink 2012), grain yield and protein content in rice (Schulthess et al. 2016), several traits in maize (Dos Santos et al. 2016) and grain yield in wheat through normalized difference vegetation index (NDVI) and canopy temperature (Sun et al. 2017). Therefore, the superiority of multi-trait models for predicting un-phenotyped individuals when using simulated data was not confirmed using experimental data.

On the other hand, simulated and empirical studies show that multi-trait models were useful for predicting traits when individuals were partially phenotyped (Rutkoski et al. 2012; Jia and Jannink 2012; Guo et al. 2014; Rutkoski et al. 2016; Hayes et al.

2017; Sun et al. 2017). Both, Rutkoski et al. 2012 and Sun et al. (2017) found advantages of multi-trait models using correlated traits from high-throughput phenotyping (i.e. NDVI and canopy temperature) in wheat. Jia and Jannink (2012) also found an improvement in predicting rust gall volume and presence or absence of rust in pine. Finally, Hayes et al. 2017 found that end-use quality traits could be better predicted using near infrared (NIR) or nuclear magnetic resonance (NMR). The results show that multi-trait models could be used to decide the optimal depth of phenotyping for each trait, mainly for expensive or difficult to measure traits (Guo et al. 2014). However, there is no evaluation on how much phenotyping of the labor intensive and expensive traits could be replaced by evaluating correlated inexpensive traits.

The objective of this work was to evaluate how multi-trait models could be used to optimize phenotyping resource allocation in breeding programs for expensive or labor intensive traits. Specifically, we evaluated whether the phenotyping of expensive and labor intensive traits could be replaced by the use of simple to measure traits without affecting the predictive ability; and we compared phenotyping strategies including the use of purposefully unbalanced designs that would require phenotyping the same number of individuals for each trait but in an unbalanced manner such that the population evaluated is significantly larger and therefore, the predictive ability potentially higher.

4.3. MATERIALS AND METHODS

4.3.1. Plant Material

Advanced inbred lines from the wheat-breeding program from the ‘Instituto Nacional de Investigación Agropecuaria’ (INIA), Uruguay, were used for this study. We used 820 advanced inbred lines for the phenotypic analysis and 1,974 advanced inbred lines for the genotypic analyses. Finally, only 495 advanced inbred lines, having both genotypic and phenotypic information were used to adjust the genomic selection models.

4.3.2. Phenotyping

The advanced inbred lines from INIA's wheat breeding program were phenotyped for eight baking quality traits. The lines were grown in 82 trials in nine location-year combinations (environment) as part of the wheat-breeding program (Table S1). There were trials from different breeding scheme stages in each environment; preliminary and advanced trials, and two maturity groups: short and long maturity. The traits were evaluated in field nurseries located in La Estanzuela (34°20'S, 57°42'W; 81 m asl), Colonia, Uruguay over 5 years (2010-2014). Additionally, approximately one third of the lines were also evaluated in Young (32°76'S, 57°57'W; 85 m asl) and Ruta 2 (33°45'S, 57°90'W; 95 m asl; Table S1). There were two to eight lines that linked the trials within and among environments (Table S2).

Eight baking quality traits were evaluated in 820 experimental advanced inbred lines. Most lines were evaluated in a single environment (~680 lines), while the most promissory lines were evaluated in multiple environments (~140 lines). Grain protein content (Pt) and test weight (TW) were determined with methods 46-12, and 55-10 of the American Association of Cereal Chemists (AACC 2000), respectively. Refined flour was obtained using the Bühler Mill (AACC Approved Method 26-21A, AACC 2000) method or equivalent. Flour attributes including wet gluten content (WG), alveograph parameters (W and L), and mixograph parameters (stability (MH) and time (MT)), were evaluated using the AACC methods 38-12, 54-30 and 54-40 (AACC 2000), respectively. Sedimentation volume (SV) was measured according to Peña et al. 1990.

4.3.3. Statistic analyses of phenotypic traits

Best linear unbiased estimation (BLUE) for bread baking quality traits were estimated from data coming from the different trials and environments using the following model:

$$[1] y_{ijk} = \mu + g_i + e_j + t_{k(j)} + \varepsilon_{ijk}$$

where y_{ijk} is the phenotypic value of the i -th genotype in j -th year-location for k -th trial, μ is the overall mean, g_i is the fixed effect of the i -th genotype, e_j is the fixed effect of j -th environment; $t_{k(j)}$ is the random effect of the k -th trial nested within j -th

environment; ε_{ijk} is the residual error for the i -th genotype in the j -th environment and k -th trial; where $t_{k(j)}$ and ε_{ijk} were random variables being $t_{k(j)} \sim N(0, \sigma_t^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma_e^2)$. The BLUEs were estimated using ‘nlme’ (Pinheiro and Bates 2017) and ‘lsmeans’ (Lenth 2016) packages from the R statistical software (R Development Core Team 2016).

To evaluate the genotype by environment interaction and the impact of unbalanced designs, Pearson’s correlation between environments was estimated using BLUEs computed by environment (as in model 1 but without the environmental effect). Additionally, genotypic, environmental, and genotype by environment interaction variance components were estimated. Finally, Pearson’s correlation and principal components analysis between traits were estimated to evaluate the correlation among traits using the BLUEs estimated from model 1. The final additive variance-covariance matrix between traits was later estimated from the genomic prediction model.

Broad sense heritability for baking quality traits was estimated following (Piepho and Möhring 2007):

$$[2] \quad H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \bar{v}/2}$$

where σ_g^2 is the genotypic variance and \bar{v} the mean variance of the difference of two adjusted means.

4.3.4. Genotyping

Leaf tissue from 1,974 lines was collected, and the CTAB method (Saghai-Maroo et al. 1984) was used to isolate DNA for genotyping-by-sequencing (Poland et al. 2012a). The TASSEL-GBS pipeline (Glaubitz et al. 2014) was run with modifications for non-reference genomes (Poland et al. 2012b). SNPs were filtered setting maximum missing value of 20%. Individuals with more than 50% missing information were also discarded. The initial calling of alleles was conducted on the large number of individuals, and then marker information for the 495 individuals with available phenotypic information was used. Therefore, the remaining 6,655 SNPs were those with minor allele frequency larger than 0.05 across the 495 individuals. SNP imputation was conducted using the multivariate normal expectation

maximization method (Endelman 2011; Poland et al. 2012a). The additive relationship matrix (K) was estimated as $K = \frac{WW'}{2\sum p_m q_m}$ where W is the centered genotypic matrix, $W_{im} = X_{im} + (1 - 2p_m)$ with X_{im} the genotype of the i-th individual for the m-th marker as $\{-1,0,1\}$ and p_m, q_m the allelic frequencies where $q_m = 1 - p_m$ (Endelman and Jannink 2012). K was estimated using the ‘rrBLUP’ package in R Statistical Software (Endelman 2011).

4.3.5. Genomic prediction models

Predictions were obtained from 495 inbred lines with phenotypic and genotypic information.

4.3.5.1. Single-trait model

Using the single-trait (ST) model, the prediction performance values were obtained for the eight baking quality traits using a Bayesian Ridge Regression (BRR) model by trait.

$$[3] y_i = 1\mu + \sum_{j=1}^{j=p} x_{ij}\beta_j + \varepsilon_i$$

where, y_i is the adjusted phenotypic mean of individual i for a single trait; μ is the overall mean; x_{ij} the score of the j-th SNP in individual i; β_j is the effect of j-th marker; and ε_i is the vector of residual errors. The conditional prior distribution is $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma^2 \sim \chi^{-2}(\sigma^2 | df, S)$ for residuals, and $\beta_j \sim N(0, \sigma_\beta^2)$ with $\sigma_\beta^2 \sim \chi^{-2}(df_\beta, S_\beta)$ for genotypic values. The total genetic values (g_i) were predicted as $g_i = \sum_{j=1}^{j=p} x_{ij}\beta_j$. Starting values were set for the degrees of freedom of the inverse-chi squared distributions (df) as 5 and scale parameters were calculated as $S = \text{var}(y) * 0.5$ following Pérez and de los Campos (2014). We used 1,500 burn-in and 3,000 iterations for the Gibbs-sampler algorithm implemented in the ‘BGLR’ package (Pérez and de los Campos 2014). Prediction accuracies for the ST model were estimated using only one cross-validation approach CV1 (de Leon et al. 2016), explained below and in Fig 1.

4.3.5.2. Multi-trait models

Multi-trait (MT) models were estimated fitting a Bayesian Multivariate Gaussian model estimating an unstructured variance-covariance matrix between traits (Σ) and a residual matrix (R). The multi-trait model is:

$$[4] y = 1\mu + Zu + \varepsilon$$

where y is a vector of $N \times t$ length (N individuals and t traits), μ is the means vector of length $N \times t$; u is a vector of predicted genetic values of the individuals for all traits with $u \sim N(0, \Sigma \otimes K)$ and ε is a vector of residuals with $\varepsilon \sim N(0, R \otimes I)$, where K is the realized additive relationship matrix among individuals estimated from the markers and, Σ and R are the variance-covariance matrices for the genetic and residual effects for each individual in all traits, respectively, estimated using a Gibbs sampler algorithm with 1,500 burn-in and 3,000 iterations. Σ was estimated as an unstructured matrix and R as a diagonal matrix, assuming heterogeneous variance and no correlation between traits, to ensure model adjustment. To estimate Σ and R Scaled-Inverse Chi-square prior distributions were assigned to $\Sigma \sim \chi^{-2}(df_{\Sigma}, S_{\Sigma})$ and $R \sim \chi^{-2}(df_R, S_R)$, with arbitrary assigned initial scale random matrix $S_{\Sigma} = I_t$, $S_R = 1_t$ and degrees of freedom $df_{\Sigma} = 4$ and $df_R = 1_t$. The predictions were obtained using the ‘MTM’ package in R (de los Campos and Grüneberg 2016). Finally, a multi-trait model without marker information was evaluated where K was an identity matrix.

4.3.6. Cross-validation scheme

Prediction accuracies were estimated using two main strategies of cross-validation also shown in Fig 1. The first cross-validation strategy (CV1 following de Leon et al. 2016) used phenotypic and genotypic information from a random set of the advanced inbred lines to train the model (for example 60% of the population or 297 individuals). Then, the remaining lines (for example 40% or 198) were predicted using genotypic data only. Pearson’s correlations between the adjusted phenotypic means (model 1) and their predicted values (model 3) were estimated. This process was iterated 100 times selecting different sets of lines each time. This scheme of cross-validation was used for ST and MT models (ST-CV1 and MT-CV1).

The second cross-validation strategy (CV2 following de Leon et al. 2016) used phenotypic and genotypic information from a random set of lines (for example 60% or

297 individuals) from the trait of interest to train the model. In addition, phenotypic and genotypic information for all lines of correlated traits was used. The trait of interest was predicted in the lines not phenotyped for the trait of interest (for example 40% or 198 individuals, Fig 1). Pearson's correlations between adjusted phenotypic means (model 1) and predicted values (model 4) were estimated. This process was iterated 100 times selecting different set of lines each time. This scheme of cross-validation was used only for the MT models (MT-CV2). Multi-trait models were used to predict traits including information from up to three correlated traits which were chosen based on their relationships revealed by principal component analyses and their Pearson's correlations.

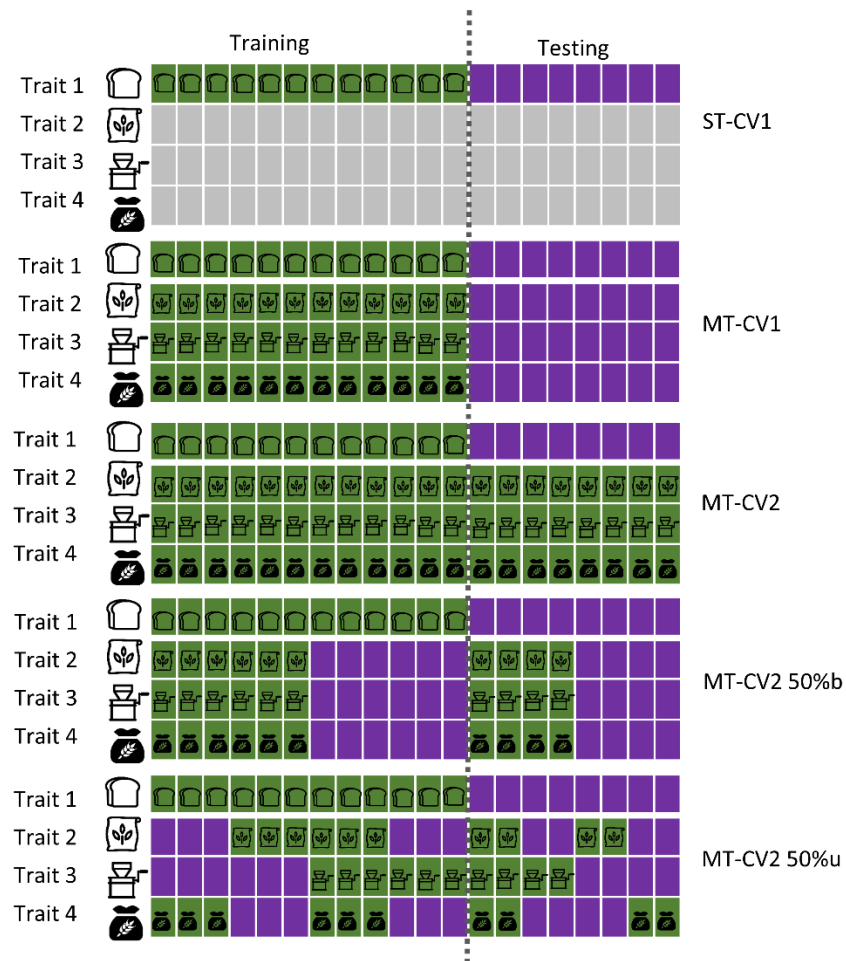
4.3.7. Improving efficiency of phenotyping

In order to evaluate the possibility to train the models with fewer phenotyped lines, we trained the MT models using 50 to 396 individuals (10 to 80%) from the training population and 100% of the lines phenotyped for the other correlated traits for W, L, and MH. This strategy was used to predict each trait at a time using three correlated traits from the same group of traits (4T), two traits from the group (3T), or one trait from the group (2T). The 2T models were constructed with SV for both W and MH, and with WG for L. The 3T models were constructed with SV and MH for W, SV and TW for MH, and WG and Pt for L. We evaluated the accuracy of the predictions using 100 iterations in all traits and CV2.

Prediction for alveograph and mixograph parameters (MH, W and L)

In order to improve the efficiency to predict complex traits such as the parameters from the alveograph or mixograph, we compared the prediction of MH, W and L using different depth of phenotypic information on correlated traits. The mixograph parameter MT was not evaluated because it is poorly correlated to other traits in its group. First, we evaluated different sizes of this training population for the predicted traits when information on the other three traits was present only in 50% of the lines in a balanced or unbalanced manner. We used the same approach as before with 10 to 80% of the training population for predicting the traits but using phenotypic information for only 50% of the lines for each correlated trait. To mask 50% of the lines we followed two strategies, designed as balanced (MT-CV2-50%b) or

unbalanced (MT-CV2-50%u). The balanced strategy masks the same 50% of the lines in all correlated traits. In the unbalanced strategy, each trait is also phenotyped in 50% of the individuals, but different individuals are phenotyped for different traits. To mask 50% of the lines, both, the training and the testing sets were divided in four equal parts. Then, two different sets from the training population and two different sets from the testing population were masked in each correlated trait. The lines were randomly assigned to each of the four sets. This procedure was conducted for 100 random iterations (Fig 1).







-  Created by T Com from Noun Project
-  Created by Mohamad Arif Prasetyo from Noun Project
-  Created by Ted Casper from Noun Project
-  Created by OCHA Visual Information Unit from Noun Project

Fig.1 Predictive strategies for an expensive or difficult to measure trait (i.e. Trait 1). ST-CV1: single-trait prediction cross validation 1 where a trait is predicted at a time, we used 60% of individuals as the training population (phenotyped and genotyped, green) and 40% of the individuals as the testing population (genotyped but not phenotyped, purple) as an example; MT-CV1: multi-trait prediction for new un-phenotyped individuals, we used 60% of individuals as the training population (phenotyped for all traits and genotyped, green) and 40% of the individuals as the testing population (genotyped but not phenotyped for any trait, purple) as an example; MT-CV2: multi-trait prediction with cross-validation 2 where 100% of the information from three correlated traits is available for the individuals to be predicted, we used 60% of individuals as the training population (phenotyped for all traits and genotyped, green) and 40% of the individuals as the testing population (phenotyped for correlated traits but not for the targeted trait, and genotyped, purple) as an example; MT-CV2-50%b: multi-trait prediction cross-validation 2 with 50% of balanced information in three correlated traits where the same 50% of individuals were phenotyped for each correlated traits; MT-CV2-50%u: multi-trait prediction cross-validation 2 with 50% of unbalanced information in three correlated traits where 50% of individuals were phenotyped for each correlated trait but in an unbalanced manner such that each individual is not necessarily phenotyped for all correlated traits. Rectangles represents lines and colors the presence (green) or absence (purple) of phenotypic information.

4.4. RESULTS

4.4.1. Phenotypic characterization of the population

Genotypic means for the multi-environment evaluation were used for genomic predictions since genotype by environment interaction among years was low (i.e. high correlation between environments, large heritability across environments, and relatively low proportion of the total phenotypic variance explained by the genotype by environment interaction; Table S2). Two groups of correlated traits were defined using principal component analysis (Fig 2) and Pearson's correlation between traits (Fig 3). The first and second principal components explained 40 and 20% of the total phenotypic variance, respectively. Both groups were represented by four traits. Group

1 includes MH, SV and W with high correlation and TW with intermediate correlation with all other traits (Fig 3). Group 2 includes Pt, WG, MT and L. MT had a low and negative correlation with all traits in this group. The heritability for each trait was medium (0.36-0.64, Fig 3).

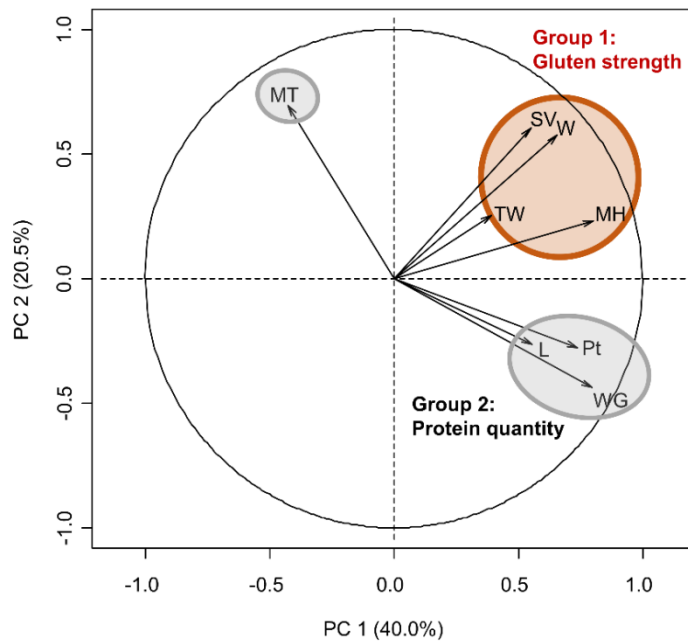


Fig.2 Principal Component analysis for the eight baking bread quality traits.

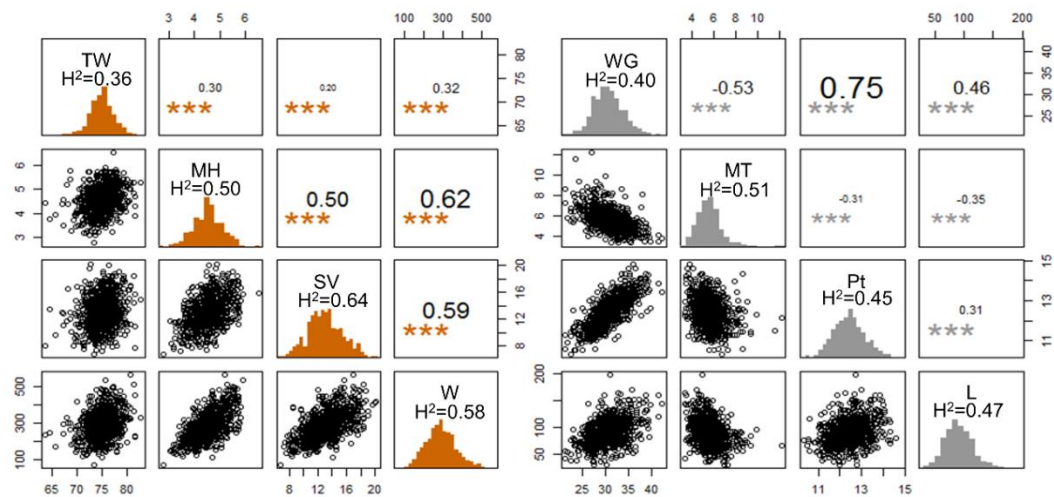


Fig.3 Scatter plot matrix with Pearson's correlations and phenotypic distributions and trait heritability (on the diagonal) for each group of traits. Group1 (left) including: TW: test weight; MH: mixograph height; SV: SDS sedimentation volume; W: alveograph parameter W. Group2 (right) including: WG: Wet gluten; MT: mixing time; Pt: grain

protein content; L: alveograph L. *** indicates significantly different from 0 with $\alpha=0.05$.

4.4.2. Multi-trait genomic predictions

Similar predictive ability of ST and MT models was found for predicting new un-phenotype individuals (ST-CV1 and MT-CV1, Fig 4). The traits with the highest predictive ability were TW and Pt ($r_{(y,\hat{y})} = 0.43$ for both), while the trait with the lowest predictive ability was L ($r_{(y,\hat{y})} = 0.24$, Fig 4).

4.4.3. Multi-trait predictions using correlated traits

Using information of correlated traits from predicted individuals increased the predictive ability for all traits (MT-CV1 vs MT-CV2, Fig 4). The improvement in predictive ability through CV2 was different for each trait and related to the correlation between traits. In group 1, TW was the trait with the smallest increase in predictive ability (Fig 4) due to its low correlation to other traits (Fig 3 Fig 3). On the other hand, the increase obtained for highly correlated traits, MT, W and SV was high. In group 2, WG was the trait with the largest increase in predictive ability, and this was the trait with the highest correlation to the others three traits in the group (Fig 3 and Fig 4).

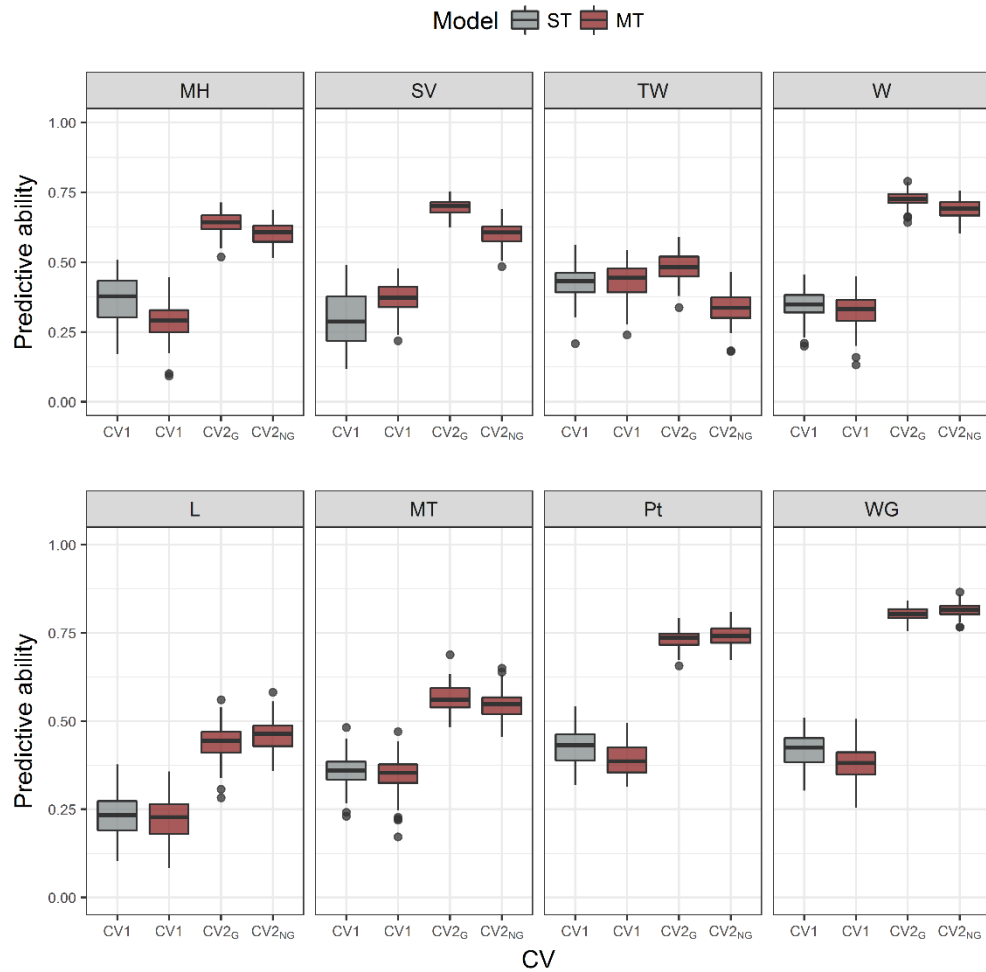


Fig.4 Predictive ability ($r(y, \hat{y})$) for eight baking bread quality traits. Single-trait prediction model (ST-CV1), and multi-trait prediction model (MT) with two schemes of cross-validation (MT-CV1, predicting new individuals and MT-CV2, predicting individuals phenotyped for correlated traits) with or without genotypic information (CV2_G and CV2_{NG}, respectively). Group1 of traits (above): TW: test weight; MH: mixograph height; SV: SDS-sedimentation volume; W: alveograph parameter W. Group 2 of traits (below): WG: wet gluten; MT: mixograph parameter-mixing time; Pt: grain protein content; L: alveograph parameter L.

4.4.4. Replace phenotyping

Using correlated traits from predicted individuals, the training population size can be reduced up to 30% of its size without significantly affecting the predictive ability of the model (Fig 5). In addition, the inclusion of marker information improved

2-14% the predictive ability of multi-trait models for W while there was no improvement for L (Fig 5).

One highly correlated trait (2T) increased more than 50% the predictive ability compared to the single trait model for both W and L (Fig 5). Two highly correlated traits (3T) increased the predictive ability 14% for W and 3% for L compared to the model with one correlated trait (Fig 5). TW did not contribute substantially to W predictions while MH and SV improved the predictive ability regardless of the model used (Fig 5). In addition, MT and Pt did not contribute to L predictions while WG increased the predictive ability for L.

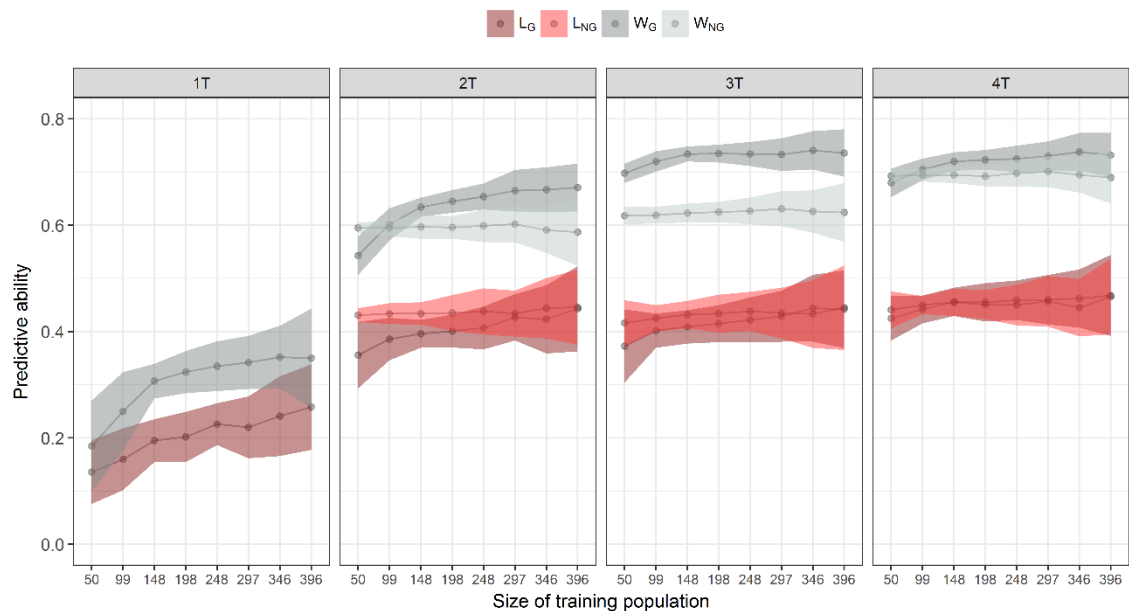


Fig.5 Predictive ability ($r(y, \hat{y})$) and standard deviation (shadowed interval) for the alveograph parameters W and L using different sizes of the training population (N=495) on the predicted trait. The traits were predicted using the multi-trait model cross-validation 2 (MT-CV2) with four, three, or two traits (4T, 3T, 2T), and single trait model (ST-CV1) to predict one trait (1T). Prediction were assessed using genotypic information (W_G and L_G) and without genotypic information (W_{NG} and L_{NG}).

To predict an expensive trait using correlated traits with equal phenotyping cost, a purposefully unbalanced phenotyping design with 50% of each trait was explored

(MT-CV2 50%u). In this case, each trait was phenotyped for 50% of the individuals but for different individuals with some overlaps. This strategy yielded higher predictive ability than using even more phenotyping but only in the training population (MT-CV1 vs MT-CV2 50%u, Fig 6). For example, 396 individuals phenotyped for W on a ST model had a predictive ability of 0.361 ± 0.09 , while phenotyping 99 individuals for W and 495 for SV had a predictive ability of 0.404 ± 0.06 (Fig 6). The predictive ability using MT-CV2 50%u for two traits was between 28% to 34% higher than the ST model for MH, W and L (Fig 6). Deep phenotyping on correlated trait always reached higher predictive ability than reducing the phenotyping to 50% of the lines for these traits. The predictions obtained using an unbalanced strategy were slightly larger than the balanced strategy. However, there were no differences in the predictive ability of both strategies used to phenotype 50% of the lines.

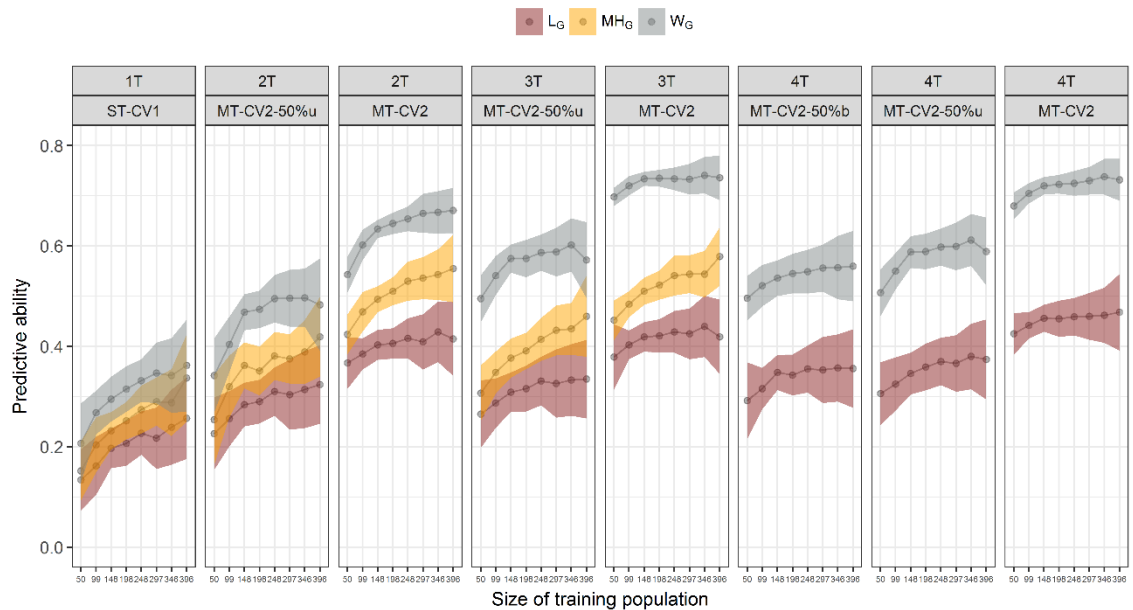


Fig.6 Predictive ability ($r(y, \hat{y})$) and standard deviation (shadowed interval) of alveograph parameters (W, L) and mixograph parameter (MH) predictions using different sizes of training population (N=495). Single-trait predictions (ST-CV1) were obtained using the predicted trait to train model (1T), or one, two, or three correlated traits to train the model (MT-CV2 for 2T, 3T and 4T, respectively). MT-CV2: multi-trait predictions with 100% of information in three correlated traits; MT-CV2-50%u:

multi-trait predictions with 50% of unbalanced information in three correlated traits; MT-CV2-50%b: multi-trait predictions with 50% of balanced information in three correlated traits.

4.5. DISCUSSION

Our grouping of traits, where one group was associated to gluten strength and the other one was related to protein quantity, is similar to that found in Vázquez et al. (2012). The high correlation we found between MH, W, and SV, was also found in others studies (Peña et al. 1994; Ruiz and Carillo 1995; Indrani et al. 2007). Model predictive ability using single-trait prediction through CV1 with 40% of randomly masked individuals (198 individuals) was lower (between 0.24-0.43) than previously found in Battenfield et al. (2016) (between 0.45 and 0.60) predicted using 20% randomly masked individuals. In addition, they were higher than those found in Hayes et al. (2017) although the strategy used to predict performance was different since we used cross-validation approaches and Hayes et al. (2017) predicted traits across years and locations.

4.5.1. Multi-trait genomic predictions

Predicting new un-phenotyped individuals is always a challenge and different strategies have been used to improve the predictive ability in those circumstances. The use of correlated trait responses has been effective when the predicted trait is of low heritability and the highly correlated trait is of high heritability (Jia and Jannink 2012; Guo et al. 2014; Jiang et al. 2015). These has been thoroughly studied both theoretically and empirically, in classic quantitative genetic studies (Falconer and Mackay 1996; Lynch and Walsh 1998) and with genomic studies (Calus and Veerkamp 2011; Jia and Jannink 2012; Guo et al. 2014; Jiang et al. 2015). However, for very complex polygenic traits, there is a small advantage of a multi-trait model with correlated responses even with high differences in heritability among traits (Jia and Jannink 2012). Furthermore, studies with real experimental data from quantitative genetics using genomic information did not show a significant improvement of multi-trait models in mice (Jiang et al. 2015), avocado (He et al. 2016), maize (Dos Santos et al. 2016), or rice (Schulthess et al. 2016). We also found that the multi-trait model

(MT-CV1) did not perform better than the single-trait model (ST-CV1). This was expected because although our traits were correlated, all traits have high heritability and because of the theoretical complexity of the traits (Nelson et al. 2006; Sun et al. 2008; Li et al. 2016).

Predictions for partially phenotyped individuals

Correlated traits can also be used to predict a correlated response when the individuals have been phenotyped for other traits (Rutkoski et al. 2012; Jia and Jannink 2012). Previous work showed high prediction accuracy using highly correlated traits, but not with intermediate to low correlated traits (Calus and Veerkamp 2011; Jia and Jannink 2012; Jiang et al. 2015). We found the same trend in our study, where the use of correlated responses using information from other traits increased the predictive ability of models, and the predictive ability was directly related with the correlation between traits. Therefore, correlated traits can be used to increase the predictive ability of the models in the lines to be predicted.

4.5.2. Predictions for replace phenotyping

We showed that the use of correlated traits from predicted individuals (MT-CV2) increase the predictive ability of the models. The next question we wanted to address was how much we could reduce the depth of phenotyping of an expensive trait (i.e. W or L in our study based on prices from the Canadian Grain Commission, Wheat Marketing Center, and AIB International), and in consequence the training population size, by using correlated traits without compromising the predictive ability of the model. It has been widely proven that smaller population sizes reduce prediction accuracy (Asoro et al. 2011; Heffner et al. 2011; Rincent et al. 2012; Akdemir et al. 2015; Rutkoski et al. 2015; Cericola et al. 2017). However, our hypothesis was that by using correlated traits we could off-set the effect of smaller population sizes. This hypothesis was tested with a range of population sizes (i.e. 50 to 396 individuals or 10 to 80%) and with real data. We found that the training population could be reduced up to 30% of the total population without significantly affecting the predictive ability of the models if correlated traits were used. Our results show that it is possible to effectively design training populations where expensive or difficult to phenotype traits are phenotyped at a smaller depth than cheaper or easier to phenotype correlated traits.

Our results were obtained with individuals from the training population chosen at random. Other studies found that optimizing the training population to select the most predictive individuals instead of using a random sample increase the predictive ability (Akdemir et al. 2015; Isidro et al. 2015; Rincent et al. 2017). We would therefore expect that our results are the baseline for the gain that could be achieved by using replaced phenotyping when the training population is optimized.

We found that the increase in predictive ability including marker information was marginal compared to the MT model using phenotypic data from correlated traits when a large number of highly correlated traits was used. However, Crain et al. (2018) showed the importance of using marker information to predict traits in a new environment where phenotypic correlation could be lower than expected due to genotype by environment interaction. We conducted cross-validation using the same population, but this is not the situation during selection in a breeding program.

We showed that expensive traits could be assessed for fewer individuals without affecting its predictive ability if information of correlated trait is used from all individuals. This requires extensive phenotyping for all correlated traits. Our results showed that using 50% of correlated information (MT-CV2-50%u) the predictive ability was lower than the full phenotyping (MT-CV2) models, but the predictive ability was still high. In addition, the use of an unbalanced strategy to reduce phenotyping on correlated traits was slightly better than reducing phenotyping using a balanced strategy (MT-CV2-50%u vs MT-CV2-50%b). Therefore, unbalanced phenotyping of correlated traits could be another approach to predict traits that are expensive to phenotype or labor intensive.

Finally, we evaluated if the inclusion of more than one trait increases the predictive ability of the model if this trait is highly correlated. We found that models with two highly correlated traits are better than models with one highly correlated trait. However the increase in predictions is low with the addition of a second correlated trait. Therefore, it will be important to evaluate the cost of prediction using two instead of one correlated trait, balancing gain in accuracy with the costs of using another trait to help predictions. The use of mildly correlated traits such as TW, MT, and Pt was not useful.

4.6. CONCLUSION

The use of MT models is useful to improve the predictive ability of partially phenotyped individuals. Expensive or difficult to phenotype traits can be phenotyped in smaller populations if the predicted individuals are phenotyped fully or partially for less expensive correlated traits. Particularly we found that the use of only one correlated trait in the model was the most effective way to increase the predictive ability with fewer resources.

Compliance with ethical standards

Conflict of interests: Authors declare they have no conflict of interest.

4.7. REFERENCES

- AACC (2000) AACC International approved methods of analysis., 11th edn. AACC International, St. Paul, MN. <http://methods.aaccnet.org/toc.aspx>. Accessed 23 Jan 2018
- Akdemir D, Sanchez JI, Jannink JL (2015) Optimization of genomic selection training populations with a genetic algorithm. *Genet Sel Evol* 47:38
- Asoro FG, Newell M a., Beavis WD, Scott MP, Jannink J-L (2011) Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. *Plant Genome J* 4:132–144
- Battenfield SD, Guzmán C, Gaynor RC, Singh RP, Peña RJ, Dreisigacker S, Fritz AK, Poland JA (2016) Genomic Selection for Processing and End-Use Quality Traits in the CIMMYT Spring Bread Wheat Breeding Program. *Plant Genome* 9:
- Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic Prediction of Breeding Values when Modeling Genotype \times Environment Interaction using Pedigree and Dense Molecular Markers. *Crop Sci* 52:707–719
- Calus MPL, Veerkamp RF (2011) Accuracy of multi-trait genomic selection using different methods. *Genet Sel Evol* 43:26
- Cericola F, Jahoor A, Orabi J, Andersen J, Janss L (2017) Optimizing Training Population Size and Genotyping Strategy for Genomic Prediction Using Association Study Results and Pedigree Information. A Case of Study in KASP. *PLoS One* 12:e0169606
- Ceron-Rojas JJ, Crossa J, Arief VN, Basford K, Rutkoski J, Jarquín D, Alvarado G, Beyene Y, Semagn K, DeLacy I (2015) A Genomic Selection Index Applied to Simulated and Real Data. *G3 (Bethesda)* 5:2155–2164
- Crain J, Mondal S, Rutkoski J, Singh RP, Poland J (2018) Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. *Plant Genome* 11:170043
- de Leon N, Jannink J, Edwards JW, Kaeppler SM (2016) Introduction to a Special Issue on Genotype by Environment Interaction. *Crop Sci* 56:2081–2089
- de los Campos G, Grüneberg A (2016) MTM Package. <http://quantgen.github.io/MTM/vignette.html>. Accessed 23 Jan 2018

- de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193:327–345
- Dos Santos JPR, De Castro Vasconcellos RC, Pires LPM, Balestre M, Von Pinho RG (2016) Inclusion of dominance effects in the multivariate GBLUP model. *PLoS One* 11:e0152045
- Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* 4:250–255
- Falconer DS, Mackay TFC (1996) *Introduction to quantitative genetics*, 4th edn. Ronald Press Company, New York
- FAO (2017) Food and Agriculture Organization of the United Nations. <http://www.fao.org/faostat/en/#compare>. Accessed 23 Jan 2018
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9:e90346
- Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G (2014) Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet* 15:30–36
- Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177:2389–2397
- Hamer RJ, MacRitchie F, Weegels PL (2009) CHAPTER 6: Structure and Functional Properties of Gluten. In: *WHEAT: Chemistry and Technology*. AACC International, Inc., 3340 Pilot Knob Road, St. Paul, Minnesota 55121, U.S.A., pp 153–178
- Hayes BJ, Walker JPC, Kant ALCS (2017) Accelerating wheat breeding for end - use quality with multi - trait genomic predictions incorporating near infrared and nuclear magnetic resonance - derived phenotypes. *Theor Appl Genet* 130:2505–2519
- Hazel LN (1943) The genetic basis for constructing selection indexes. *Genetics* 28:476–490
- He D, Kuhn D, Parida L (2016) Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction. *Bioinformatics* 32:i37–i43

- Heffner EL, Jannink J, Sorrells ME (2011) Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. *Plant Genome* 4:65–75
- Henderson CR, Quaas RL (1976) Multiple Trait Evaluation Using Relatives' Records. *J Anim Sci* 43:1188–1197
- Heslot N, Jannink J-L, Sorrells ME (2015) Perspectives for Genomic Selection Applications and Research in Plants. *Crop Sci* 55:1–12
- Indrani D, Manohar RS, Rajiv J, Rao GV (2007) Alveograph as a tool to assess the quality characteristics of wheat flour for parotta making. *J Food Eng* 78:1202–1206
- Isidro J, Jannink J-L, Akdemir D, Poland J, Heslot N, Sorrells ME (2015) Training set optimization under population structure in genomic selection. *Theor Appl Genet* 128:145–158
- Jia Y, Jannink J-L (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 192:1513–1522
- Jiang J, Zhang Q, Ma L, Li J, Wang Z, Liu J (2015) Joint prediction of multiple quantitative traits using a Bayesian multivariate antedependence model. *Heredity (Edinb)* 115:29–36
- Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–756
- Lenth R V (2016) Least-Squares Means: The R Package lsmeans. *J Stat Softw* 69:1–33
- Li C, Bai G, Chao S, Carver B, Wang Z (2016) Single nucleotide polymorphisms linked to quantitative trait loci for grain quality traits in wheat. *Crop J* 4:1–11
- Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink JL (2011) Genomic Selection in Plant Breeding. *Knowledge and Prospects. Adv Agron* 110:77–123
- Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland, MA, USA
- MacRitchie F (1992) Physicochemical Properties of Wheat Proteins in Relation to Functionality. *Adv Food Nutr Res* 36:1–87

- Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Nelson JC, Andreescu C, Breseghello F, Finney PL, Daisy G, Perretant MR, Leroy P, Bergman CJ, Pe RJ, Qualset CO, Sorrells ME (2006) Quantitative trait locus analysis of wheat quality traits. *Euphytica* 149:145–159
- Peña RJ, Amaya A, Rajaram S (1990) Variation in Quality Characteristics Associated with Some Spring 1B / IR Translocation Wheats. *J Cereal Sci* 12:105–112
- Peña RJ, Zarco-Hernandez J, Amaya-Celis A, Mujeeb-Kazi A (1994) Relationship Between Crhromosome 1B-encoded Glutenin Subunit Composition and Bread-making Quality Characteristic of some Durum Wheat (*Triticum turgidum*) Cultivars. *J Cere* 19:243–249
- Pérez P, de los Campos G (2014) Genome-Wide Regression and Prediction with the. *Genet Soc Am* 198:483–495
- Piepho H-P, Möhring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* 177:1881–1888
- Pinheiro J, Bates D (2017) Linear and Nonlinear Mixed Effects Models. <https://cran.r-project.org/web/packages/nlme/nlme.pdf>. Accessed 23 Jan 2018
- Poland J a, Brown PJ, Sorrells ME, 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:e32253
- Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink J-L (2012b) Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *Plant Genome J* 5:103–113
- R Development Core Team (2016) R: The R Project for Statistical Computing. <https://www.r-project.org/>. Accessed 23 Jan 2018
- Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodríguez VM, Moreno-Gonzalez J, Melchinger A, Bauer E, Schoen C-C, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, Moreau L (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics* 192:715–728

- Rincent R, Oury EKHMFX, Rousset M, Allard V (2017) Optimization of multi - environment trials for genomic selection based on crop models. *Theor Appl Genet* 130:1735–1752
- Ruiz M, Carillo M (1995) Relationships between different prolamin proteins and some quality properties in durum wheat. *Plant Breed* 114:40–45
- Rutkoski J, Benson J, Jia Y, Brown-guedira G, Jannink J, Sorrells M (2012) Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat. *Plant Genome* 5:51–61
- Rutkoski J, Poland J, Mondal S, Autrique E, Pérez LG, Crossa J, Reynolds M, Singh R (2016) Canopy Temperature and Vegetation Indices from High-Throughput Phenotyping Improve Accuracy of Pedigree and Genomic Selection for Grain Yield in Wheat. *G3 (Bethesda)* 6:2799–2808
- Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL, Sorrells ME (2015) Efficient Use of Historical Data for Genomic Selection: A Case Study of Stem Rust Resistance in Wheat. *Plant Genome* 8:
- Saghai-Marooif M a, Soliman KM, Jorgensen R a, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. *Proc Natl Acad Sci U S A* 81:8014–8018
- Schulthess AW, Yu W, Miedaner T, Wilde P, Reif JC, Zhao Y (2016) Multiple- trait- and selection indices- genomic predictions for grain yield and protein content in rye for feeding purposes. *Theor Appl Genet* 129:273–287
- Shewry PR, Hey SJ (2015) The contribution of wheat to human diet and health. *Food Energy Secur* 4:178–202
- Smith HF (1936) A Discriminant Function for Plant Selection. *Ann Eugen* 7:240–250
- Sun H, Lu J, Fan Y, Zhao Y, Kong F, Li R, Wang H, Li S (2008) Quantitative trait loci (QTLs) for quality traits related to protein and starch in wheat. *Prog Nat Sci* 18:825–831
- Sun J, Rutkoski JE, Poland JA, Crossa J, Jannink J, Sorrells ME (2017) Multitrait , Random Regression , or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. *Plant Genome* 10:

- VanRaden PM (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* 91:4414–4423
- Vázquez D (2009) Aptitud industrial de trigo. In: Inst. Nac. Investig. Agropecu. http://www.inia.uy/Publicaciones/Documentos_compartidos/18429130709133540.pdf. Accessed 24 Jan 2018
- Vázquez D, Berger AG, Cuniberti M, Bainotti C, Zavariz de Miranda M, Scheeren PL, Jobet C, Zúñiga J, Cabrera G, Verges R, Peña RJ (2012) Influence of cultivar and environment on quality of Latin American wheats. *J Cereal Sci* 56:196–203
- Williams RMA, Brien LOB, Eagles HAC, Solah VAA, Jayasena VA (2008) The influences of genotype , environment , and genotype x environment interaction on wheat quality. *Aust J Agric Res* 59:95–111

5. COMPARACIÓN DE ESTRATEGIAS DE PREDICCIÓN DE CRUZAMIENTOS PARA LA SELECCIÓN DE LOS MEJORES CRUZAMIENTOS EN UN PROGRAMA DE MEJORAMIENTO GENÉTICO DE TRIGO

5.1. RESUMEN DEL CAPÍTULO

La selección de cruzamientos dentro de un programa de mejoramiento es un paso clave en la mejora genética del cultivo ya que impacta directamente en los resultados obtenidos. Sin embargo, es una etapa en la cual la disponibilidad de información de sustento para esta decisión es escasa. Existen muchos posibles cruzamientos y sólo unos pocos son factibles de realizar, dado que cada cruzamiento incrementa de manera exponencial el número de líneas que se evaluarán en las distintas etapas del programa. El incremento de información genómica disponible y el desarrollo de modelos de predicción hacen que hoy en día existan más elementos disponibles de apoyo a la decisión sobre los cruzamientos a realizar. El objetivo de este trabajo fue evaluar y comparar la selección de cruzamientos considerando la media de la progenie o considerando la media del 10% superior de la progenie, prediciendo la varianza genética asumiendo equilibrio de ligamiento (V_{LE}) en la población y considerando el desequilibrio de ligamiento (V_{LD}). Se evaluaron 1.465 líneas provenientes del programa de mejoramiento de INIA-Uruguay para rendimiento de grano, con un total de 3.884 SNPs identificados. Para calidad panadera, se evaluaron tres caracteres en 5.984 líneas del programa de mejoramiento de CIMMYT-México con 1.164 SNPs identificados. La media de la progenie se predijo como la media del desempeño de ambos padres, que se estimó utilizando el modelo RR-BLUP y Bayes Lasso incluyendo la información genotípica. La varianza de los padres se estimó considerando los marcadores en equilibrio de ligamiento según Endelman (2011) y considerándolos en desequilibrio de ligamiento según Mohammadi et al. (2015). Los resultados muestran que para rendimiento no hubo diferencias entre los métodos de selección, por lo que la selección de cruas utilizando las medias de los padres es la mejor estrategia. Sin embargo, para los caracteres de calidad, considerar la varianza genética permitió seleccionar cruas que no fueron seleccionadas considerando sólo

1el desempeño de ambos padres. Por lo que en este caso, el considerar la varianza fue importante, aunque no hubo diferencias entre los dos métodos usados para la predicción de la varianza (V_{LE} y V_{LD}).

Strategies for Selecting Crosses Using Genomic Prediction in Two Wheat Breeding Programs

Bettina Lado, Sarah Battenfield, Carlos Guzmán, Martín Quincke, Ravi P. Singh, Susanne Dreisigacker, R. Javier Peña, Allan Fritz, Paula Silva, Jesse Poland,* and Lucía Gutiérrez*

Abstract

The single most important decision in plant breeding programs is the selection of appropriate crosses. The ideal cross would provide superior predicted progeny performance and enough diversity to maintain genetic gain. The aim of this study was to compare the best crosses predicted using combinations of mid-parent value and variance prediction accounting for linkage disequilibrium (V_{LD}) or assuming linkage equilibrium (V_{LE}). After predicting the mean and the variance of each cross, we selected crosses based on mid-parent value, the top 10% of the progeny, and weighted mean and variance within progenies for grain yield, grain protein content, mixing time, and loaf volume in two applied wheat (*Triticum aestivum* L.) breeding programs: Instituto Nacional de Investigación Agropecuaria (INIA) Uruguay and CIMMYT Mexico. Although the variance of the progeny is important to increase the chances of finding superior individuals from transgressive segregation, we observed that the mid-parent values of the crosses drove the genetic gain but the variance of the progeny had a small impact on genetic gain for grain yield. However, the relative importance of the variance of the progeny was larger for quality traits. Overall, the genomic resources and the statistical models are now available to plant breeders to predict both the performance of breeding lines per se as well as the value of progeny from any potential crosses.

Core Ideas

- Cross prediction strategies for grain yield and baking quality traits were compared.
- Crosses for all parent combinations were obtained via genomic prediction models.
- Mid-parent selection was similar to accounting for variance when selecting yield.
- The variance had a larger impact in cross predictions for quality traits.

THE MAIN OBJECTIVE of plant breeding is to increase the yield, productivity, adaptation, and quality of crops while optimizing resource use (Allard 1960). Genetic gain in plant breeding is accomplished through the selection of best genetic combinations between

B. Lado, L. Gutiérrez, Statistics Dep., Facultad de Agronomía, Univ. de la República, Garzón 780, Montevideo 12900, Uruguay; L. Gutiérrez, Dep. of Agronomy, Univ. of Wisconsin–Madison, 1575 Linden Dr, Madison, WI 53706; S. Battenfield, J. Poland, Wheat Genetics Resource Center, Dep. of Plant Pathology, 1712 Claflin Rd., Kansas State Univ., Manhattan, KS 66506; S. Battenfield (current address), AgriPro Wheat, Syngenta, 11783 Ascher Rd, Junction City, KS, 66441; C. Guzmán, R.P. Singh, S. Dreisigacker, R.J. Peña, CIMMYT, El Batán, Mexico, Mexico; A. Fritz, Dep. of Agronomy, Kansas State Univ., Manhattan, KS 66506; M. Quincke, P. Silva, Programa Nacional de Investigación Cultivos de Secano, Instituto Nacional de Investigación Agropecuaria, Est. Exp. La Estanzuela, Colonia 70000, Uruguay. B. Lado and S. Battenfield contributed equally. Received 14 Dec. 2016. Accepted 18 Mar. 2017.
*Corresponding authors (jpoland@ksu.edu; gutierrezcha@wisc.edu). Assigned to Associate Editor Nicholas Tinker.

Published in Plant Genome
Volume 10. doi: 10.3835/plantgenome2016.12.0128

© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA
This is an open access article distributed under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Abbreviations: BL, Bayes Lasso; GS, genomic selection; INIA, Instituto Nacional de Investigación Agropecuaria; RIL, recombinant inbred line; RR-BLUP, ridge regression best linear unbiased predictor; QTL, quantitative trait loci; SNP, single nucleotide polymorphism; V_{LD} , variance accounting for linkage disequilibrium; V_{LE} , variance assuming linkage equilibrium.

genotypes (Fehr, 1987). The most important traits selected in breeding are generally complex quantitative traits that are influenced by many genes and the environment (Falconer and Mackay, 1996; Bernardo, 2010). Historically, strategies for selecting superior individuals have relied on phenotypic evaluation of individuals and pedigree information. However, at the end of the twentieth century, marker-assisted selection strategies were sought. To this end, quantitative trait loci (QTL) were detected to identify genes involved in the expression of complex traits through biparental QTL mapping (Lander and Botstein, 1989; Lande and Thompson, 1990), or genome-wide association studies (Jannink et al., 2001). These genetic markers were then used in marker-assisted selection strategies (Dekkers and Hospital, 2002). However, not all of the segregating QTL are identified in a given study and the ones that are identified are largely overestimated (Beavis, 1994). Therefore, methodologies for QTL identification have been insufficient to capture the effects of the many genes involved in common complex traits (Manolio et al., 2009) and the actual use of marker-assisted selection for identified QTL has been very limited in breeding programs (Bernardo, 2008).

Genomic prediction (Meuwissen et al., 2001) provides an alternative method of using genetic information in breeding decisions. Rather than using individual QTL information, genomic prediction involves the prediction of genotypic performance accounting for all genome-wide marker effects simultaneously (Meuwissen et al., 2001). When these predictions are used as a selection strategy, it is called genomic selection (GS). Genomic selection has been extensively researched in plants (reviewed in Heslot et al., 2015) while the implementation in plant breeding programs has been evaluated through optimization of the training population (Asoro et al., 2011; Zhao et al., 2013; Liu et al., 2015; Isidro et al., 2015) and genotype \times environment interaction (Burgueño et al., 2012; Heslot et al., 2014; Jarquín et al., 2014; Lado et al., 2016). Additionally, strategies of GS to select the best crosses by modeling the variance of a cross have been proposed but no study has assessed the consequences in cross selection (Bernardo, 2014; Mohammadi et al., 2015; Tiede et al., 2015).

Crossing is one of the main decisions of a breeding program. The number of possible crosses is orders of magnitude larger than the number of feasible crosses (Witcombe et al., 2013). Although breeders attempt to take into consideration all available information on the potential parents to determine which crosses to make, many crosses are discarded in subsequent years, as they do not deliver superior progeny (Koeberner and Summers, 2003; Heslot et al., 2015). Methods to select optimum crosses accounting for the performance per se of each line (e.g., yield, adaptability, quality) were initially proposed on the basis of the progeny's mean and variance. The variance was predicted using morphological data, pedigree information, a few molecular markers, or a combination of these (Souza and Sorrells, 1991; Melchinger et al.,

1998; Utz et al., 2001; Bertan et al., 2007; Casassola and Brammer, 2013). These methods were then extended to genome-wide marker coverage (Endelman, 2011; Bernardo, 2014; Mohammadi et al., 2015; Tiede et al., 2015). The progeny mean was well-predicted from the mid-parent value (Bernardo, 2014). However, genetic variances were difficult to predict using either phenotypic, pedigree, or genetic distances (Souza and Sorrells, 1991; Melchinger et al., 1998; Utz et al., 2001; Hung et al., 2012) but could be adequately modeled by simulating progeny by using genomic prediction (Bernardo, 2014; Mohammadi et al., 2015; Tiede et al., 2015). Endelman (2011) proposed a strategy of estimating the mean and variance for the progeny of a cross using genomic data, where the progeny mean was estimated as the mid-parent value of the estimated parental performance. In this method, the variance was estimated from the marker effects, assuming linkage equilibrium (V_{LE}) between markers. Another strategy used to estimate progeny variance with large genomic data is to simulate the progeny, accounting for the true linkage structure and recombination, and then to predict the performance of the progeny (Bernardo, 2014; Mohammadi et al., 2015; Tiede et al., 2015).

The aim of this research was to evaluate the selection of crosses based on progeny variance, predicted either by assuming linkage equilibrium (V_{LE}) or by accounting for linkage disequilibrium (V_{LD}) between the parents. Specifically, we evaluated the impact of cross selection based only on mid-parent performance or mid-parent performance in a combination with progeny variance, where the variance was estimated either by assuming V_{LE} or V_{LD} . We analyzed data for grain yield in the INIA spring bread wheat breeding program in Uruguay, and grain protein content, mixograph optimal mixing time, and loaf volume quality traits from the CIMMYT spring bread wheat breeding program in Mexico. We evaluated multiple traits and breeding programs to limit possible trait or program bias in the prediction methodology results.

Materials and Methods

Plant Material

INIA

A total of 1465 spring bread wheat lines from the INIA Wheat Breeding Program were used as the training population for GS for grain yield. The INIA lines consisted of all the lines from the preliminary yield trails (F_7) from 2010, 2011, and 2013, as well as the advanced (F_8) and elite (F_9) yield trials from 2010.

Grain yield evaluations were conducted in 35 environments in Uruguay including five locations evaluated in 4 yr and one location with four sowing dates. Locations used to evaluate the genotypes were Dolores (33°50'S, 58°14'W; 15 m asl), Durazno (33°33'S, 56°31'W; 91 m asl), La Estanzuela (34°20'S, 57°42'W; 81 m asl), Young (32°76'S, 57°57'W; 85 m asl), and Ruta2 (33°45'S, 57°90'W; 95 m asl). Four sowing dates were evaluated in

La Estanzuela (LE1, LE2, LE3, and LE4). The evaluation years were 2010 to 2014. The trials from 2012 and most of the locations of 2014 were not included in this study because of a strong genotype by environment interaction (see experimental details in Lado et al., 2016).

CIMMYT

In total, 6095 spring bread wheat lines from the CIMMYT Global Bread Wheat Breeding Program were evaluated for quality traits (Battenfield et al., 2016). The CIMMYT lines were materials in the preliminary yield trials between 2009 and 2015. All lines were grown in Ciudad Obregon, Sonora, Mexico (27°29'N, 109°56'W; 40 m asl). Three phenotypes were chosen to represent different stages of testing wheat as grain, dough, and final product. Grain protein content was determined by near-infrared spectroscopy, using NIR Systems 6500 (Foss, Denmark) using AACC International (2000a) Method 39–10, and reported at 12.5% moisture basis. Grain samples were tempered and milled using Brabender Quadrumat Senior (C. W. Brabender OHG, Germany). Mixograph optimal mixing time was assessed using the Swanson and Working Mixograph (National MFG Co., National Manufacturing Company, Lincoln, NE) according to AACC International Method 54–40A (AACC International, 2000b). Loaf volume was assessed on pup loaves baked according to AACC International (2000c) Method 10–09. Loaf volume and mixing time tests used the adjustment for optimal water absorption described in Guzmán et al. (2015).

Genotyping

Tissue was collected and the CTAB method (Saghai-Marouf et al., 1984) was used to isolate DNA for the genotyping-by-sequencing protocol as described in Poland et al. (2012a). The TASSEL-GBS pipeline (Glaubitz et al., 2014) was run with modifications for nonreference genomes in the INIA material (Poland et al., 2012b). The TASSEL Version 5.2 with alignment to the *T. aestivum* International Wheat Genome Sequencing Consortium genome assembly version 2.25 (Mayer et al., 2014) using Bowtie2 version 2.24 (Langmead and Salzberg, 2012) was used in the CIMMYT lines. Genotyping-by-sequencing tags with single nucleotide polymorphisms (SNPs) were mapped against the POPSEQ Synthetic W7984 by Opata M85 reference to determine the map position (Chapman et al., 2015). Single nucleotide polymorphisms were filtered by setting maximum missing values of 20% for the INIA and 50% for the CIMMYT populations. Different thresholds for missing data were used to keep the balance between coverage and good SNP quality at each population. Low tolerance for missing values was used to increase the reliability of the imputed SNPs. Individuals with more than 50% missing information were discarded. The final number of individuals were 1465 and 5984 in the INIA and CIMMYT populations, respectively. SNP imputation was conducted using multivariate normal expectation maximization

imputation (Endelman, 2011; Poland et al., 2012a). This method assumes that marker genotypes follow a multivariate normal distribution. The imputation was done through multiple regression starting with the mean of the marker and was then updated using maximum likelihood to estimate the mean and covariance between lines with nonmissing data.

Phenotypic Data Analysis

Phenotypic best linear unbiased estimations were obtained for the grain yield of all genotypes present in each trial. Field analysis was conducted according to an experimental design that consisted of a series of smaller α -design trials connected through common checks. The following model was used to estimate INIA grain yield genotypic means:

$$y_{ijklm} = \mu + g_i + e_j + t_{k(i)} + r_{l(jk)} + b_{m(ljk)} + \varepsilon_{ijklm}, \quad [1]$$

where μ is the overall mean, g_i is the fixed effect of the i^{th} genotype; e_j is the fixed effect of j^{th} year–location combination; $t_{k(i)}$ is the random effect of the k^{th} trial nested within the j^{th} year–location; $r_{l(jk)}$ is the random effect of the l^{th} replicate nested within the k^{th} trial and j^{th} year–location; $b_{m(ljk)}$ is the random effect of the m^{th} incomplete block nested within the l^{th} replicate, k^{th} trial, and the j^{th} year–location; and ε_{ijklm} is the residual error for the i^{th} genotype in the m^{th} block within the l^{th} replicate in the k^{th} trial located in the j^{th} year–location; and $t_{k(i)}$, $r_{l(jk)}$, $b_{m(ljk)}$, and ε_{ijklm} are random variables: $t_{k(i)} \sim N(0, \sigma_t^2)$, $r_{l(jk)} \sim N(0, \sigma_r^2)$, $b_{m(ljk)} \sim N(0, \sigma_b^2)$, and $\varepsilon_{ijklm} \sim N(0, \sigma_e^2)$. The best linear unbiased estimations were estimated with the nlme package (Pinheiro et al., 2013) in R statistical software (R Development Core Team, 2016). Heritability was estimated via the model shown in Eq. [1] but with genotypes as the random effect, being $g_i \sim N(0, G\sigma_g^2)$. G is the genotypic variance–covariance matrix between lines estimated by using markers, as was proposed by VanRaden (2008), and those estimated by using the A.mat function in the rrBLUP package (Endelman, 2011). Genotypic variance (σ_g^2) and the average pairwise variance error of the best linear unbiased predictor (\bar{v}_{BLUP}) were estimated using the sommer R package (Covarrubias-Pazarán, 2016). Heritability for grain yield was estimated following Cullis et al. (2006):

$$H^2 = 1 - (\bar{v}_{BLUP} / 2\sigma_g^2). \quad [2]$$

For the CIMMYT population, quality data were not replicated within or across years and no statistical design was used within years, since only high-yielding selections from the yield trial were advanced to quality testing. The heritability of the quality traits was estimated as in Battenfield et al. (2016) for CIMMYT data from the ridge regression best linear unbiased predictor (RR-BLUP) model, which only models additive effects. Briefly,

$$H^2 = (\sigma_p^2 - \sigma_e^2) / \sigma_p^2, \quad [3]$$

where σ_p^2 is the phenotypic variance of the trait in the training population and σ_e^2 is the error variance of the RR-BLUP model. The σ_e^2 is assumed to include genotypic variance caused by dominance and epistatic effects, as well as any nongenetic errors. This leaves only additive genetic variance when σ_e^2 is subtracted from σ_p^2 .

Genomic Prediction Models

Two genomic prediction models were compared in order to estimate marker effects and the performance of each parent. The ridge regression model was evaluated via the rrBLUP package (Endelman, 2011) and the Bayesian Lasso (BL) model was calculated via the BGLR package (Pérez and de los Campos, 2010), both in R software (R Development Core Team, 2016). Genomic prediction models incorporated the genotypic matrix \mathbf{X} codified as $(-1, 0, 1)$, where +1 indicates that loci were homozygous to the most frequent allele at a given locus, -1 indicates that loci were homozygous to the least frequent allele at given locus, and 0 indicates heterozygous loci. The missing SNPs were imputed as fractions between -1 and 1, and then were rounded to the nearest whole number (to 1 if the SNP value was ≥ 0.5 , -1 if the SNP value was less than or below -0.5 , and 0 if the SNP value was between -0.5 and 0.5). For BL, initial arbitrary values of trait heritability and error variance were used, then values were optimized via prediction accuracy to calculate the hyperparameter rate and shape from the double exponential distribution. Next, the marker effects and parent performance values were estimated from the posterior distribution. For the ridge regression model, marker effects were estimated by using mixed models with markers considered as random with $N(0, I\sigma_{\text{mrk}}^2)$. After that, to predict the performance of the parents, mixed models with lines (g) as random effects were used with $N(0, G\sigma_g^2)$, where G is the variance-covariance matrix estimated as in VanRaden (2008) via the A.mat function in the rrBLUP package (Endelman, 2011). The models were selected because of their use of different statistical approaches (mixed models and Bayesian, respectively) and their different assumptions about marker effects on predictions.

Random cross-validation was conducted by using 60% of the lines to train the model, predicting the remaining 40% of entries for both the INIA and CIMMYT datasets separately. Each model was iterated 100 times and then, the model with the best predictive ability was used for subsequent analysis. All genotyped parents were included in the training population to estimate the marker effects and the parents' performance for grain yield ($n = 1465$), grain protein content, mixing time, and loaf volume ($n = 5984$), via the model with the best predictive ability. However, as there were no differences in the accuracy of predictions between the models, the RR-BLUP model was used to estimate marker effects for all traits for the remaining analyses.

Cross Selection Based on the Performance of Simulated Progeny

Progeny of all possible pairwise cross combinations were simulated and their performance was predicted for grain yield in the INIA population. However, in the CIMMYT population, only the lines evaluated in 2015 ($n = 1425$) were used as parents for progeny simulation and prediction. Reciprocal crosses were not predicted, as negligible maternal effects were assumed to be present (McNeal et al., 1968; Jumbo and Carena, 2008).

The V_{LE} of the progeny within each cross ($V_{LE(i)}$) was predicted, taking into account allelic differences between the parents and markers effects for each trait, following Endelman (2011):

$$V_{LE(i)} = \sum_{k=1}^{k=M} (1 - (p_{k+} - p_{k-})^2) \times \hat{u}[k]^2, \quad [4]$$

where p_{k+} is the frequency of the k^{th} SNP of biallele +1 for parents of progeny in i^{th} cross and p_{k-} is the same for the biallele -1, and \mathbf{u} is the vector of marker effects with a length M (the number of SNPs). Marker effects were estimated via RR-BLUP. For example, if both parents had the allele +1, the frequencies were $p_{k+} = 1$ and $p_{k-} = 0$; if one parent had the allele +1 and the other had -1, the frequencies were $p_{k+} = 0.5$ and $p_{k-} = 0.5$; if both parents had the alleles 0, $p_{k+} = 0$ and $p_{k-} = 0$, and so on. Next, the distribution of predicted performance values for the population of progeny in each cross was simulated, assuming a normal distribution for the progeny. The parameters used in the normal distribution were the mean (equal to the mid-parent value) and the variance (equal to the progeny variance estimated as in Eq. [4]). The mean of the top 10% of the progeny was estimated for each cross.

To predict V_{LD} , we used the PopVar package described in Mohammadi et al. (2015) and implemented in R software (R Development Core Team, 2016). Within PopVar, recombinant inbred lines (RILs) were simulated via the Rqtl package (Broman et al., 2003), which requires ordered markers with genetic distances to account for recombination. The function simulates recombination points along the chromosome by using independent crossovers, following the Stahl model (Stahl, 1979). Using this package, we simulated 1000 RILs for each cross and then estimated the predicted performance of each individual simulated RIL. The mean performance of the 1000 RILs was used to estimate the progeny mean performance of each cross. The variance of the predicted performance of the 1000 RILs was used to estimate the progeny variance performance of each cross. The mean of the top 10% of the progeny was estimated for each cross.

Finally, the best 1000 crosses were selected for grain yield, grain protein, and loaf volume on the basis of either their mid-parent performance value or the predicted performance of the top 10% of their progeny, assuming V_{LD} or V_{LE} . For mixing time, where an intermediate optimal time is desired, the best crosses were selected on the basis of having the lowest variance within a window of targeted predicted progeny performance values (i.e., 2.5–4.0 min.).

Weighting Progeny Variance in Cross Selection

Five values of performance were established as thresholds to illustrate the trade-offs of selection on the basis of maximizing the expected mean or accounting for the variance. Therefore, five groups with different weights on the relative importance of the variance were established. The maximum parental predicted performance minus 20, 40, 60, 80 or 100% of the difference between the maximum and the mean predicted performance was used to define the groups G1 (20%), G2 (40%), G3 (60%), G4 (80%), and G5 (100%). One hundred crosses were selected within each group, taking into account the mean predicted progeny performance above the threshold and the maximum variance for three traits (mixing time was excluded from this analysis because the selection criteria for this trait are different). Finally, the mean of all progenies and the mean of the top 10% of the progeny for the 100 selected progenies were calculated.

Results and Discussion

Genotype Data

For INIA data, a total of 81,044 SNPs were identified after the TASSEL pipeline was run and duplicate markers were removed. Of these, 23,771 were mapped using the POPSEQ genetic framework. Finally, we kept 3884 SNPs with less than 20% missing data and 1465 individuals with less than 50% missing data. For CIMMYT data, a total of 31,166 SNPs were identified after the TASSEL pipeline was run and duplicate markers were removed. Of these, 8443 SNPs were mapped using POPSEQ. Finally, we kept 1164 SNPs with less than 50% missing data and 5984 individuals with less than 50% missing data. The SNPs were distributed in all chromosomes with a median of 0 and 0 cM and maximum gap of 61 cM and 94 cM, for the INIA and CIMMYT genetic maps, respectively (Supplemental Fig. S1).

Phenotype Data

Mean grain yield for the INIA dataset was 5160 kg ha⁻¹ (SE = 853) and the heritability was 0.67. The means of the three end-use quality traits from CIMMYT were 11.97% (SE = 0.77) for grain protein content, 3.13 min (SE = 0.80) for mixing time, and 802 cm³ (SE = 59) for loaf volume. The heritability of the CIMMYT quality traits were 0.67, 0.71, and 0.68 for grain protein content, mixing time, and loaf volume, respectively.

Genomic Selection Models

High correlations between observed and predicted breeding values were found for grain yield in the INIA dataset ($r_{(\hat{y}, \bar{y})} = 0.42$) and for grain protein content ($r_{(\hat{y}, \bar{y})} = 0.63$), mixing time ($r_{(\hat{y}, \bar{y})} = 0.64$), and loaf volume ($r_{(\hat{y}, \bar{y})} = 0.59$) in the CIMMYT dataset. The predictive ability using both models, BL and RR-BLUP, were similar. As expected, the correlation of marker effects between models decreased at the extreme values for each trait (Supplemental Fig. S2). The BL method induces less shrinkage

than RR-BLUP in marker effects estimation as it sets some of the marker effects to zero and thus BL estimates fewer but larger effects than RR-BLUP (Park and Casella, 2008). Mixing time shows the largest shrinkage effect for RR-BLUP (Supplemental Fig. S2). This was expected, because mixing time is influenced by a few genes with a large effect (Payne et al., 1987). However, when the marker effect values were combined to predict performance of the parents, there was no difference between GS models (Supplemental Fig. S2). It has been documented elsewhere that different GS models had similar prediction ability in plants (Heslot et al., 2012). Thus because of the similarity in accuracies between RR-BLUP and BL predictions, all further analyses were conducted using RR-BLUP model.

Progeny Prediction

The predicted mean performance of the progeny was perfectly correlated between the mid-parent value and mean of the simulated progenies for all traits (data not shown). This result was expected because in an additive model, the expected value of the mean of the progeny calculated as the mid-parent value is the same as the mean of the RILs' performance. The same results were found in maize (*Zea mays* L.) for silking date and protein (Bernardo, 2014) and in barley (*Hordeum vulgare* L.) for yield and deoxynivalenol content (Mohammadi et al., 2015).

The correlation between V_{LD} and V_{LE} was different for each trait and was probably associated with trait complexity (Fig. 1). The order of the variance correlation between estimations was grain yield < grain protein content < loaf volume < mixing time. Variance predictions of more complex traits are more inconsistent (Bernardo, 2008) and thus modeling variances for these traits has less impact in progeny predictions. This is probably because the error in the estimates of marker effects has a stronger influence on the variance than on the mean predictions of a cross (Zhong and Jannink, 2007). The higher variances correlation for mixing time found in our study can be explained by the fact that mixing time is a trait that is influenced by a few major genes (Payne et al., 1987; Simons et al., 2012).

The variance of predicted performance from 1000 progeny was larger for V_{LD} than for V_{LE} (Fig. 2 and Fig. 3). A potential explanation for this is as follows. The genetic variance for each pair of markers is the sum of the variance of each marker, plus twice the covariance between the two markers (Falconer and Mackay, 1996). When we assume linkage equilibrium, the markers were independent and therefore, the covariance between the markers was zero. If markers are in the coupling phase, the covariance is positive and therefore V_{LE} is smaller than V_{LD} (Lynch and Walsh, 1998; Bernardo, 2010). However, if the markers are in the repulsion phase, the covariance between markers becomes negative and therefore, V_{LE} is larger than V_{LD} (Lynch and Walsh, 1998; Bernardo, 2010). Hence, the relative magnitude of the variances (i.e., V_{LE} and V_{LD}) will depend on the marker

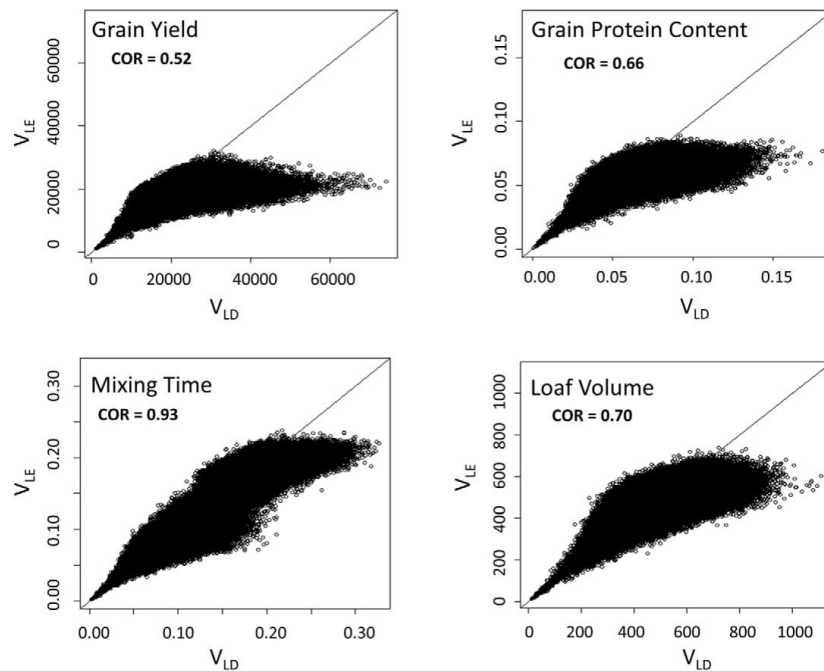


Fig. 1. Correlation between the variance of predicted progeny performance accounting for linkage disequilibrium (V_{LD}) or assuming linkage equilibrium (V_{LE}), for wheat grain yield, grain protein content, mixing time, and loaf volume.

phasing. In our case, we believe we have coupling phases in several situations. For example, one of the crosses that exhibits the largest differences in variance estimated by V_{LD} and V_{LE} showed a predominance of coupling phases for most chromosomes (Fig. 4).

We found a triangular relationship between variance and the mean of predicted progeny performance for the crosses in all traits when modeling V_{LD} , but not when modeling V_{LE} (Fig. 2 and Fig. 3). This relationship was stronger for grain yield than for the other traits (Fig. 2 and Fig. 3). This triangular relationship is generated when lines with similar trait values at the end of their distribution (i.e., low \times low or high \times high) are crossed, resulting in progeny with extreme values for mean and low variance (Bernardo, 2014). On the other hand, when lines from opposite extremes are crossed (i.e., low \times high), the progeny have intermediate means with higher variance. This is supported by previous findings by Bernardo (2014) and Mohammadi et al. (2015). Furthermore, as in Bernardo (2014), we found a linear relationship between the mean and variance of crosses from the same parent prediction when using V_{LD} (Fig. 4), which caused the triangular shape. Recombination between extreme parents would account for differences arising from coupling and repulsion linkage. These differences may not have an immediate impact but could become relevant for longer-term selection (Bernardo, 2014).

Cross Selection

When selecting crosses based on mid-parent values or on the basis of the mean of the top 10% of the progeny via two variance estimates (V_{LD} and V_{LE}), we observed that many of the selected crosses and parents were the same for all three selection methods (Fig. 2 and Fig. 3). Of the 1000 best crosses based on the two variance methods, there were 879 crosses in common for grain yield, 757 for grain protein content, 769 for mixing time, and 738 for loaf volume (Fig. 2B,E and Fig. 3B,E). In addition, 837 of the common crosses selected for grain yield were the same ones with the highest mid-parent values. This was also confirmed by the high correlation ($r = 0.94$, $p < 0.001$) between total progeny mean and the mean of the top 10% of the predicted progeny performance for the 1000 best crosses (selected by top 10%). These results indicate that the predicted mean progeny performance is the strongest driver for selecting superior crosses for grain yield. Thus regardless of the approach for predicting progeny variance, the crosses with the highest mid-parent values also provided the best top 10% of the progeny. This was largely independent of the genetic distance between parents. Mid-parent strategies are superior for selecting crosses than a combination of mid-parent and progeny variance when there is more variation in the progeny means than in the progeny variance (Zhong and Jannink, 2007).

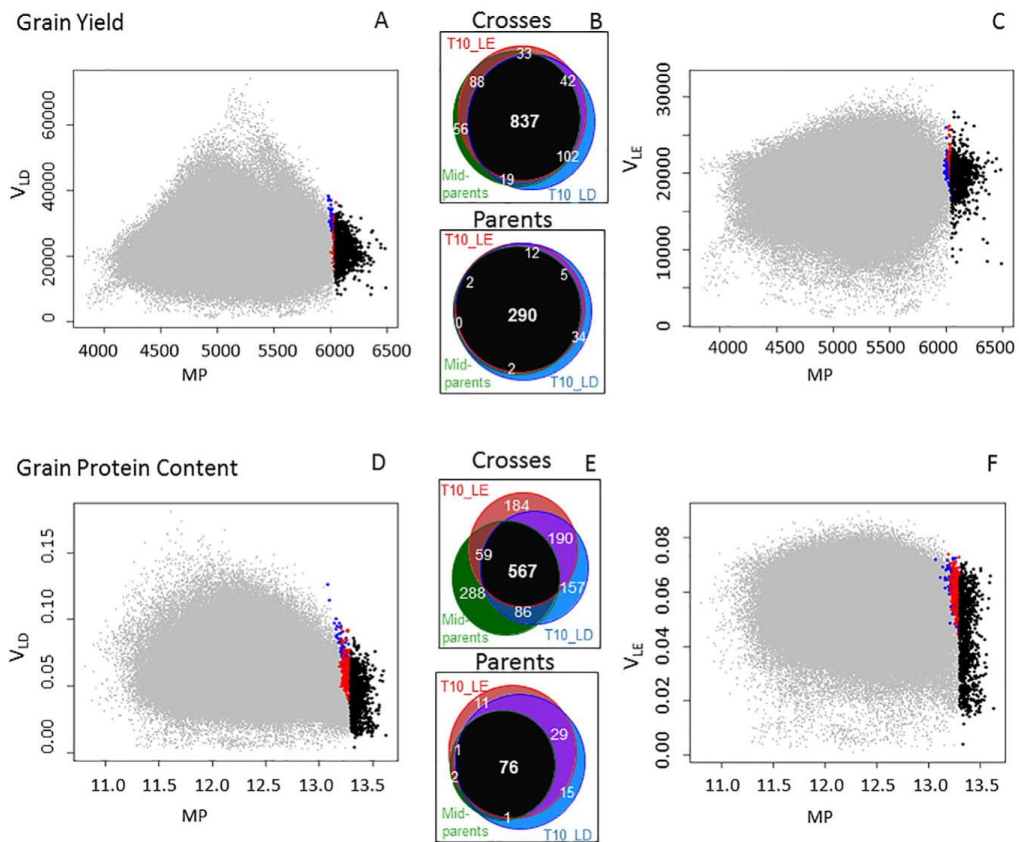


Fig. 2. Cross performance characterization. Expected mid-parent performance (MP) versus the predicted progeny variance accounting for linkage disequilibrium (V_{LD}) for yield (A) and grain protein content (D). Expected MP versus the predicted progeny variance assuming linkage equilibrium (V_{LE}) for yield (C) and grain protein content (F). The top 1000 crosses selected by MP (green), V_{LD} (blue), V_{LE} (red), and crosses in common among the two groups (black) are highlighted in each plot (A, C, D, F). Venn diagrams represent the 1000 crosses; the parents of those crosses selected by using the MP (green), V_{LD} (blue), V_{LE} (red); or common crosses among the three groups (black) for grain yield (B) and grain protein content (E).

For quality traits, the percentage of crosses in common between the three cross prediction strategies were 56.7% for grain protein content and 59.8% loaf volume (Fig. 2 and Fig. 3). In addition, the correlations between the total progeny mean and the mean of the top 10% of the predicted progeny performance for the best 1000 crosses (selected by top 10%) were, 0.67 ($p < 0.001$) for grain protein content and 0.80 ($p < 0.001$) for loaf volume. These results indicate that for quality traits, the influence of variance was a more relevant factor in determining cross selection. The results found by Bernardo (2014) for silking date and protein concentration in maize are more similar to those found for wheat quality traits in which crosses with an undesirable mean value but high variance could reach the same mean as the top 10% of the progeny.

Accounting for more variance (i.e., by having lower thresholds) decreased the number of shared crosses

between V_{LD} and V_{LE} (Fig. 5, left and right plots). Additionally, the mean of the top 10% of the progeny increased as the threshold value increased, regardless of cross variance (Fig. 5, middle plots). This effect was more pronounced for grain yield, which had the highest ratio between the variance of mean predicted progeny performance values and the variance of progeny SD among crosses (Table 1).

Although the differences between the variance predicted by the strategies considering V_{LD} or V_{LE} are high for some traits, the parents and the crosses selected on the basis of the top 10% performance of the progeny are similar. Therefore, the use of V_{LE} as a substitute for V_{LD} is a reasonable compromise if computation time or the lack of a linkage map are important considerations. The use of V_{LE} could be favored when one considers only one generation of improvement. For grain yield, where all three

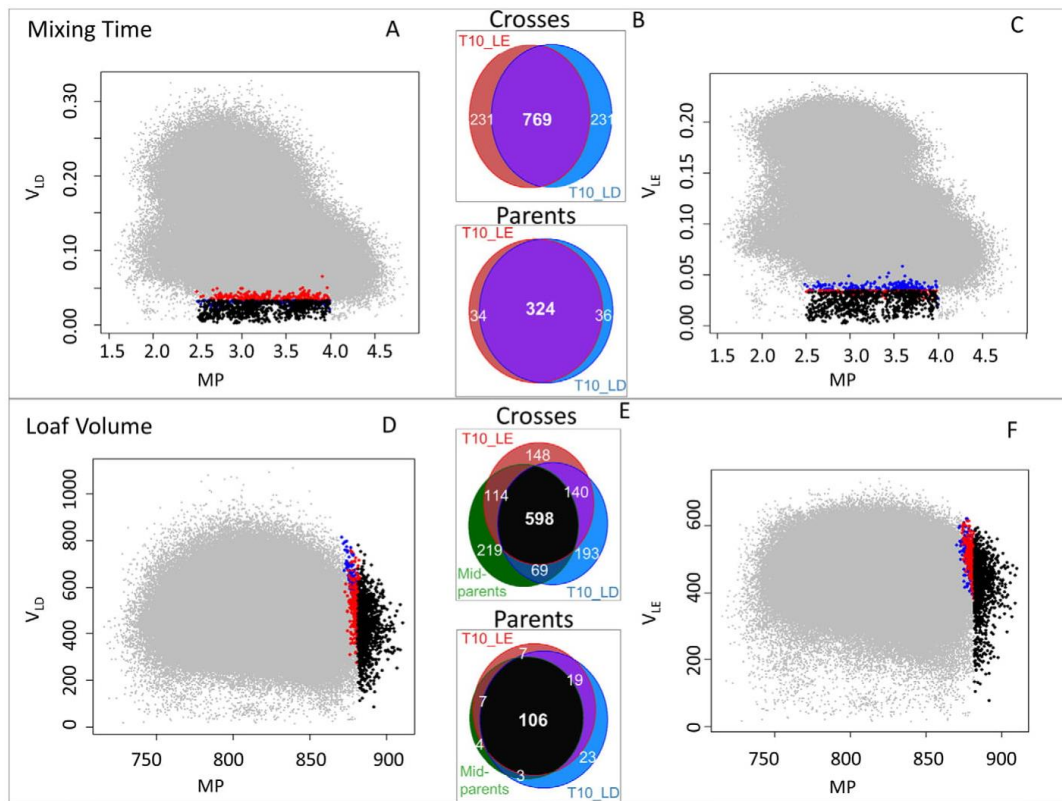


Fig. 3. Cross performance characterization. Expected mid-parent performance (MP) versus the predicted progeny variance accounting for linkage disequilibrium (V_{LD}) for mixing time (A) and loaf volume (D). Expected MP versus the predicted progeny variance assuming linkage equilibrium (V_{LE}) for mixing time (C) and loaf volume (F). The top 1000 crosses selected by using MP (green), V_{LD} (blue), V_{LE} (red), and crosses in common among the two groups (black) are highlighted in each plot (A, C, D, F). Venn diagrams represent the 1000 crosses or the parents of those crosses selected only by using V_{LD} (blue), V_{LE} (red) for mixing time (B), and MP (green), alongside common crosses among the three groups (black) for loaf volume (E).

methods gave similar results, the use of the mid-parent value would be the simplest method of cross prediction.

On the basis of our results, the best strategy of maintaining diversity is not clear. Fu (2015) discussed the difficulty of measuring and accounting for the impacts of plant breeding on crop genetic diversity. Jacobson et al. (2015) and Rutkoski et al. (2015) found that the losses of genetic diversity incurred when using genome-wide selection is higher than that for phenotypic selection, although there are situations in which some losses of diversity are necessary to obtain increased genetic gains. However, to sustain genetic gain in wheat breeding programs, genetic diversity is essential because it allows effective recombination, which, in turn, creates more diversity (Fehr, 1987). This diversity is fundamental when new challenges are encountered, such as changing races in biotic stressors, or changing environmental factors and abiotic stressors (Dreisigacker et al., 2005; Borlaug, 2007; Babiker et al.,

2015). Thus, Fu (2015) emphasized the importance of maintaining allele variability to have the possibility to improve under unfavorable conditions.

Conclusions

Mid-parent value has historically been used for cross selection in plant breeding. Here, progeny mean was estimated using genomic predicted mid-parent value, which was identical to the mean of the simulated progeny. Modeling the variance to predict the best crosses had a small impact for complex traits like grain yield. However, for quality traits, the influence of variance was a large factor in determining cross selection. Differences in selection assuming either V_{LD} or V_{LE} were small for all traits and the assumption of V_{LE} was less computationally intensive. Although selection of parental combinations based on the expected progeny variance for yield

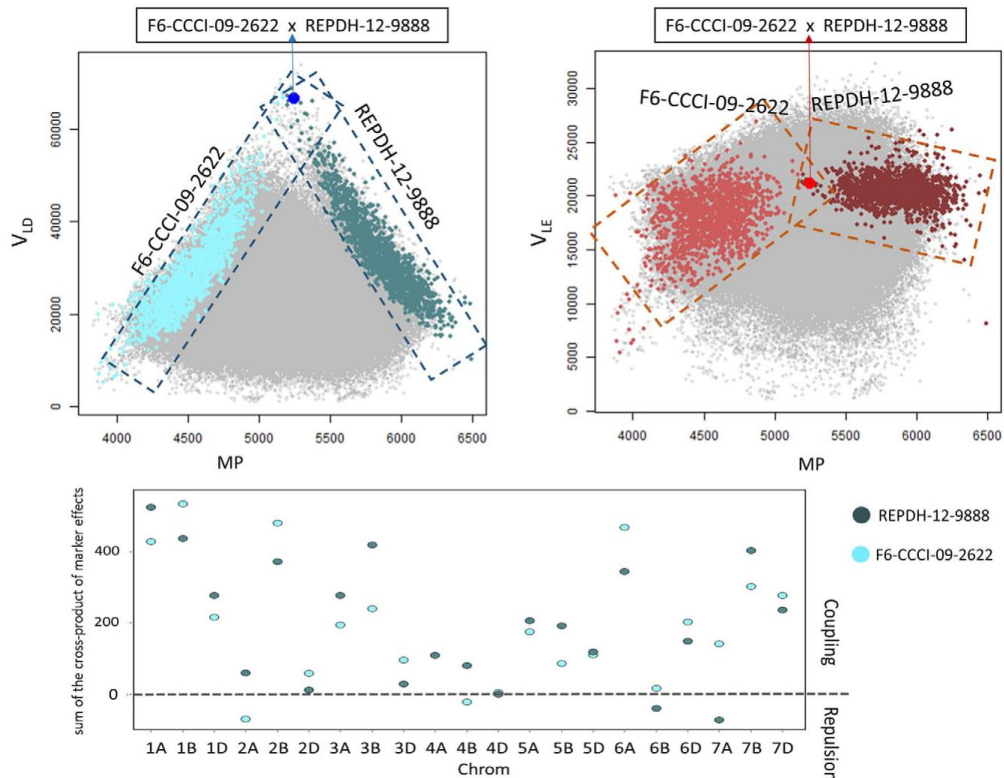


Fig. 4. Relationship between mid-parent performance (MP) and variance for all possible crosses predicted considering linkage disequilibrium (V_{LD} ; left) or linkage equilibrium (V_{LE} ; right). An extreme cross between a high-performing parent and a low-performing parent is highlighted, as well as the progeny of each of those parents (i.e., dark colors for the high-performing parent and light colors for the low-performing parent). The bottom panel of the figure represents the average allelic phasing for each chromosome calculated for adjacent close markers.

had little impact on improving the overall cross performance, maintaining genetic diversity should also be considered. In reality, this may be accomplished through the choice of parents that achieve high average yield through slightly different strategies or yield components.

Supplemental Information

Supplemental Fig. S1. Single nucleotide polymorphism (SNP) map with marker distribution across the wheat genome in the INIA (on the left) and CIMMYT (on the right) populations. Single nucleotide polymorphisms were mapped against the POPSEQ Synthetic W7984 by Opata M85 reference.

Supplemental Fig. S2. RR-BLUP vs BL scatterplots for marker effects (upper row), mid-parent value (MP, middle row), and progeny variance assuming linkage equilibrium (V_{LE} , bottom row) for grain yield, grain protein content, mixing time and loaf volume (traits by column).

Conflict of Interest Disclosure

The authors declare no conflict of interest.

Acknowledgments

Support for grain yield phenotyping was provided by INIA. We express our appreciation for the effort of the technical personnel of the INIA wheat breeding program: Leonardo Hernández, María Ferreira, Dumas Laun, Abel López, and José Flores. Support for phenotyping of the CIMMYT quality traits was provided by CGIAR Comite Regional Permanente (Permanent Regional Committee) WHEAT, Durable Rust Resistance Project, and Fondo Sectorial Secretaria de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación (Secretary of Agriculture, Livestock, Rural Development, Fishery and Food, of Mexico) Consejo Nacional de Ciencia y Tecnología (National Council on Science and Technology, of Mexico) (No. 146788: "Sistema de mejoramiento genético para generar variedades resistentes a royas, de alto rendimiento y alta calidad para una producción sustentable en México de trigo") of the Mexican government. Funding for CIMMYT GS was provided by the US Agency for International Development (USAID) Feed the Future Initiative (USAID Cooperative Agreement No. AID-OAA-A-13-0005) and the Bill & Melinda Gates Foundation through a grant to Cornell University for "Genomic Selection: The next frontier for rapid gains in maize and wheat improvement". The opinions expressed herein are those of the author(s) and do

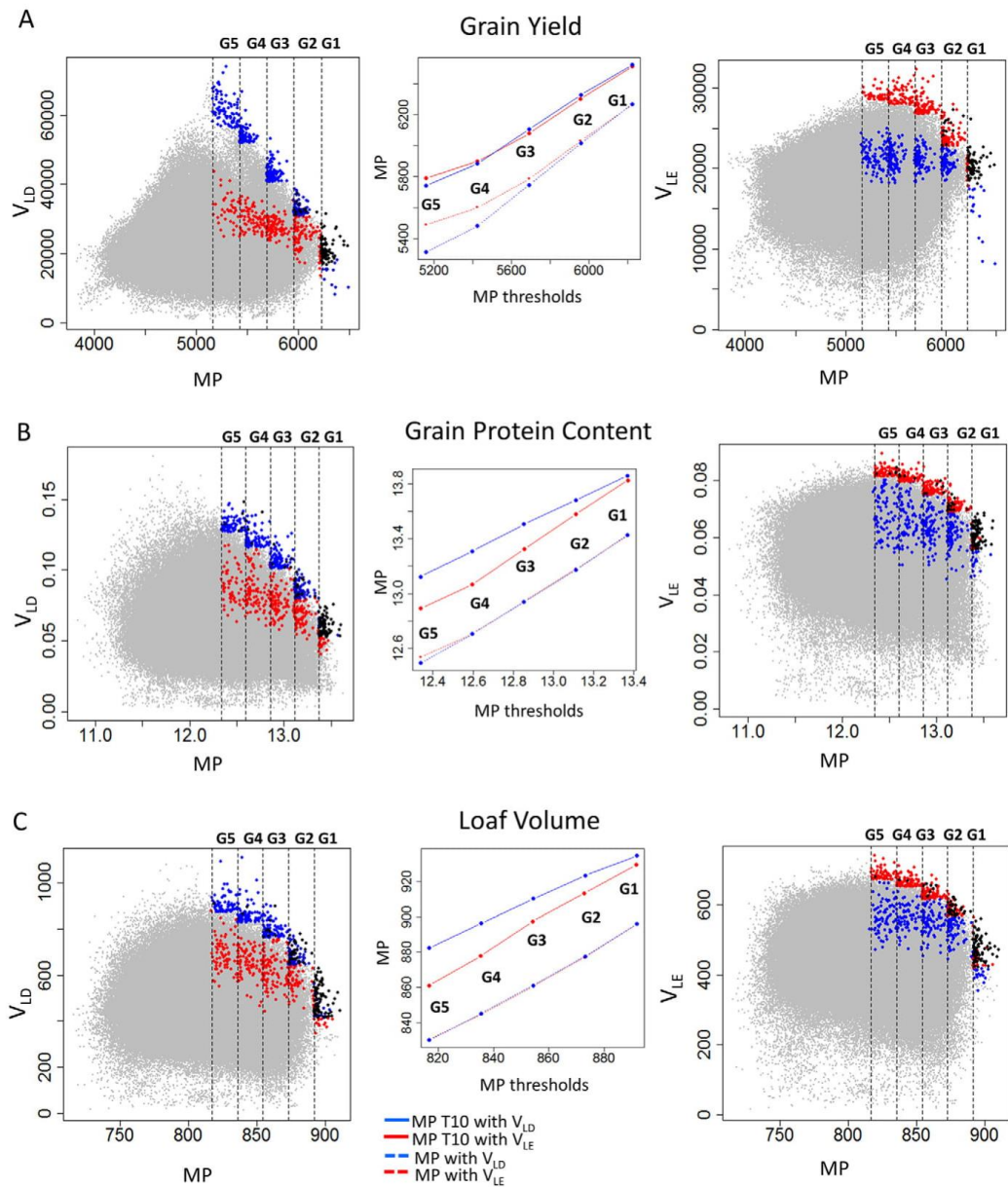


Fig. 5. The best 100 crosses selected using five progeny mean predicted performance thresholds [G1 (20%) > G2 (40%) > G3 (60%) > G4 (80%) > G5(100%)] and maximum variance are indicated in two correlation plots within each trait. Plots on the left are the expected mid-parent performance (MP) versus the variance accounting for linkage disequilibrium (V_{LD}). Plots on the right are the MP versus variance assuming linkage equilibrium (V_{LE}). Blue and red points indicate the best 100 crosses selected by accounting for linkage disequilibrium or assuming linkage equilibrium, respectively. Black points indicate the common crosses between V_{LD} and V_{LE} . The mean predicted progeny performance values of 100% (MP; broken line) or 10% (MP T10; continuous line) of the top performing progeny considering linkage disequilibrium (V_{LD} ; blue) or linkage equilibrium (V_{LE} ; red) for each selection group [G1–G5] are shown in the middle plots.

Table 1. Variance of mean predicted progeny performance ($Var(MP)$) among crosses and variance of the SD of predicted progeny performance ($\sqrt{V_{ID}}$) and the ratio between them [$Var(\sqrt{V_{ID}})/Var(MP)$]. Predictions were obtained by considering linkage disequilibrium for four traits: grain yield, grain protein content, mixing time, and loaf volume.

Trait	$Var(MP)$	$Var(\sqrt{V_{ID}})$	Ratio
Grain yield (kg ha ⁻¹)	90,969	28,400,576	312
Grain protein content (%)	0.056	0.022	2.5
Mixing time (min.)	0.070	0.012	5.8
Loaf volume (cm ³)	0.093	0.011	8.5

not necessarily reflect the views of USAID. Support for the doctoral work of SB was provided by Monsanto's Beachell-Borlaug International Scholar's Program. Support for the doctoral work of BL was provided by Agencia Nacional de Investigación e Innovación–Uruguay, through Grant POS_NAC_2013_1_11261 and by Comisión Sectorial de Investigación Científica–Uruguay, through grants in the program internships abroad.

References

AACC International. 2000a. AACCI Method 39-10.01. Near-infrared reflectance method for protein determination in small grains. AACC International. <http://methods.aaccnet.org/summaries/39-10-01.aspx> (accessed 10 May 2017).

AACC International. 2000b. AACCI Method 54-40.02. Physical dough tests. Mixograph method. AACC International. <http://methods.aaccnet.org/summaries/54-40-02.aspx> (accessed 10 May 2017).

AACC International. 2000c. AACCI Method 10-09.01. Baking quality. Basic straight-dough bread-baking method—long fermentation. AACC International. <http://methods.aaccnet.org/summaries/10-09-01.aspx> (accessed 10 May 2017).

Allard, R.W. 1960. Principles of plant breeding. John Wiley and Sons Inc., New York.

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, and J.-L. Jannink. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *Plant Genome* 4(2):132–144. doi:10.3835/plantgenome2011.02.0007

Babiker, E.M., T.C. Gordon, S. Chao, M. Newcomb, M.N. Rouse, Y. Jin, et al. 2015. Mapping resistance to the Ug99 race group of the stem rust pathogen in a spring wheat landrace. *Theor. Appl. Genet.* 128(4):605–612. doi:10.1007/s00122-015-2456-6

Battenfield, S.D., C. Guzmán, R.C. Gaynor, R.P. Singh, R.J. Peña, S. Dreisigacker, et al. 2016. Genomic selection for processing and end-use quality traits in the CIMMYT Spring Bread Wheat Breeding Program. *Plant Genome* 9(2):1–12. doi:10.3835/plantgenome2016.01.0005

Beavis, W.D. 1994. The power and deceit of QTL experiments: Lessons from comparative QTL studies. In D.B. Wilkinson, editor. *Proceedings of the 49th Annual Corn and Sorghum Research Conference*. American Seed Trade Association, Washington, DC. p. 250–265

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Sci.* 48(5):1649–1664. doi:10.2135/cropsci2008.03.0131

Bernardo, R. 2010. *Breeding for quantitative traits in plants*, 2nd ed.: Stemma Press, Woodbury, MN.

Bernardo, R. 2014. Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. *Crop Sci.* 54(1):1–33.

Bertan, I., F.I.F. De Carvalho, and A.A.C. De Oliveira. 2007. Parental selection strategies in plant breeding programs. *J. Crop Sci. Biotechnol.* 10(4):211–222.

Borlaug, N.E. 2007. Sixty-two years of fighting hunger: Personal recollections. *Euphytica* 157(3):287–297. doi:10.1007/s10681-007-9480-9

Broman, K.W., H. Wu, S. Sen, and G.A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19(7):889–890. doi:10.1093/bioinformatics/btg112

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa. 2012. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52(2):707–719. doi:10.2135/cropsci2011.06.0299

Casassola, A., and S. Brammer. 2013. Parental selection of wheat lines based on phenotypic characterization and genetic diversity. *Crop Breed. Appl. Biotechnol.* (13):49–58. doi:10.1590/S1984-70332013000100006

Chapman, J.A., M. Mascher, A. Buluç, K. Barry, E. Georganas, A. Session, et al. 2015. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol.* 16(1):26. doi:10.1186/s13059-015-0582-8

Covarrubias-Pazarán, G. 2016. Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS ONE* 11(6):E0156744. doi:10.1371/journal.pone.0156744

Cullis, B.R., B. Smith, and N.E. Coombes. 2006. On the design of early generation variety trials with correlated data. *J. Agric. Biol. Environ. Stat.* 11(4):381–393. doi:10.1198/108571106X154443

Dekkers, J.C.M., and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. *Nat. Rev. Genet.* 3(1):22–32. doi:10.1038/nrg701

Dreisigacker, S., P. Zhang, M.L. Warburton, B. Skovmand, D. Hoisington, and E. Melchinger. 2005. Genetic diversity among and within CIMMYT wheat landrace accessions investigated with SSRs and implications for plant genetic resources management. *Crop Sci.* 45(2):653–661. doi:10.2135/cropsci2005.0653

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4(3):250–255. doi:10.3835/plantgenome2011.08.0024

Falconer, D.S., and T.F. Mackay. 1996. *Introduction to quantitative genetics*. Prentice Hall, London.

Fehr, W.R. 1987. *Principles of cultivar development*. Macmillan Inc., New York.

Fu, Y.B. 2015. Understanding crop genetic diversity under modern plant breeding. *Theor. Appl. Genet.* 128(11):2131–2142. doi:10.1007/s00122-015-2585-y

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, et al. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9(2):E90346. doi:10.1371/journal.pone.0090346

Guzmán, C., G. Posadas-Romano, N. Hernández-Espinosa, A. Morales-Dorantes, and R.J. Peña. 2015. A new standard water absorption criteria based on solvent retention capacity (SRC) to determine dough mixing properties, viscoelasticity, and bread-making quality. *J. Cereal Sci.* 66:59–65. doi:10.1016/j.jcs.2015.10.009

Heslot, N., D. Akdemir, M.E. Sorrells, and J.-L. Jannink. 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127(2):463–480. doi:10.1007/s00122-013-2231-5

Heslot, N., J.-L. Jannink, and M.E. Sorrells. 2015. Perspectives for genomic selection applications and research in plants. *Crop Sci.* 55: 1–12. doi:10.2135/cropsci2014.03.0249

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink. 2012. Genomic selection in plant breeding: A comparison of models. *Crop Sci.* 52(1): 146–160. doi:10.2135/cropsci2011.06.0297

Hung, H.-Y., C. Browne, K. Guill, N. Coles, M. Eller, A. Garcia, et al. 2012. The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity* 108(5):490–499. doi:10.1038/hdy.2011.103

Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot, and M.E. Sorrells. 2015. Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* 128:145–158. doi:10.1007/s00122-014-2418-4

- Jacobson, A., L. Lian, S. Zhong, and R. Bernardo. 2015. Minimal loss of genetic diversity after genomewide selection within biparental maize populations. *Crop Sci.* 55(2):783–789. doi:10.2135/cropsci2014.10.0744
- Jannink, J.-L., M.C.A.M. Bink, and R.C. Jansen. 2001. Using complex plant pedigrees to map valuable genes. *Trends Plant Sci.* 6:337–342. doi:10.1016/S1360-1385(01)02017-9
- Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Gaucourt, J. Lorgeou, et al. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127(3):595–607. doi:10.1007/s00122-013-2243-1
- Jumbo, M.B., and M.J. Carena. 2008. Combining ability, maternal, and reciprocal effects of elite early-maturing maize population hybrids. *Euphytica* 16:325–333. doi:10.1007/s10681-007-9618-9
- Koebner, R.M.D., and R.W. Summers. 2003. 21st Century wheat breeding: Plot selection or plate detection? *Trends Biotechnol.* 21(2):59–63. doi:10.1016/S0167-7799(02)00036-7
- Lado, B., P.G. Barrios, M. Quincke, P. Silva, and L. Gutiérrez. 2016. Modeling genotype \times environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Sci.* 56:1–15. doi:10.2135/cropsci2015.04.0207
- Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124(3):743–756.
- Lander, E.S., and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121(1):185–199.
- Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9(4):357–359. doi:10.1038/nmeth.1923
- Liu, H., H. Zhou, Y. Wu, X. Li, J. Zhao, T. Zuo, et al. 2015. The impact of genetic relationship and linkage disequilibrium on genomic selection. *PLoS ONE* 10(7):1–13.
- Lynch, M., and B. Walsh. 1998. *Genetics and analysis of quantitative traits*. Sinauer Associates Inc., Sunderland, MA.
- Manolio, T., F. Collins, and N. Cox. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753. doi:10.1038/nature08494
- Mayer, K.F.X., J. Rogers, J. Dole, C. Pozniak, K. Eversole, C. Feuillet, et al. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345(6194):1251788–1251788.
- McNeal, F.H., M.A. Berg, and C.A. Watson. 1968. Reciprocal crosses and their influence on wheat quality. *Crop Sci.* 8:485–487. doi:10.2135/cropsci1968.0011183X000800040027x
- Melchinger, A.E., R.K. Gumber, R.B. Leipert, M. Vuylsteke, and M. Kuiper. 1998. Prediction of testcross means and variances among F_2 progenies of F_1 crosses from testcross means and genetic distances of their parents in maize. *Theor. Appl. Genet.* 96(3–4):503–512. doi:10.1007/s001220050767
- Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4):1819–1829.
- Mohammadi, M., T. Tiede, and K.P. Smith. 2015. PopVar: A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. *Crop Sci.* 55(5):2068–2077. doi:10.2135/cropsci2015.01.0030
- Park, T., and G. Casella. 2008. The Bayesian Lasso. *J. Am. Stat. Assoc.* 103(482):681–686. doi:10.1198/016214508000000337
- Payne, P.I., M.A. Nightingale, A.F. Krattiger, and L.M. Holt. 1987. The relationship between HMW glutenin subunit composition and the bread-making quality of British-grown wheat varieties. *J. Sci. Food Agric.* 40(1):51–65. doi:10.1002/jsfa.2740400108
- Pérez, P., and G. de los Campos. 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *Plant Genome* 3(2):106–116. doi:10.3835/plantgenome2010.04.0005
- Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team. 2013. *Linear and nonlinear mixed effects models*. Springer-Verlag: New York.
- Poland, J., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 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. doi:10.1371/journal.pone.0032253
- Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, et al. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5(3):103–113. doi:10.3835/plantgenome2012.06.0006
- R Development Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing. <http://www.R-project.org> (accessed 11 May 2017).
- Rutkoski, J., R.P. Singh, J. Huerta-Espino, S. Bhavani, J. Poland, J.L. Jannink, et al. 2015. Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. *Plant Genome* 8(2):1–10. doi:10.3835/plantgenome2014.10.0074
- Saghai-Maroo, M., K.M. Soliman, R. Jorgensen, and R.W. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci. USA* 81(24):8014–8018. doi:10.1073/pnas.81.24.8014
- Simons, K., J.A. Anderson, M. Mergoum, J.D. Faris, D.L. Klindworth, S.S. Xu, et al. 2012. Genetic mapping analysis of bread-making quality traits in spring wheat. *Crop Sci.* 52(5):2182–2197. doi:10.2135/cropsci2012.03.0175
- Souza, E., and M.E. Sorrells. 1991. Prediction of progeny variation in oat from parental genetic relationships. *Theor. Appl. Genet.* 82(2):233–241. doi:10.1007/BF00226219
- Stahl, F.W. 1979. Special sites in generalized recombination. *Annu. Rev. Genet.* 13(5):7–24. doi:10.1146/annurev.ge.13.120179.000255
- Tiede, T., L. Kumar, M. Mohammadi, and K.P. Smith. 2015. Predicting genetic variance in bi-parental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. *Mol. Breed.* 35(10):1–13. doi:10.1007/s11032-015-0390-6
- Utz, H.F., M. Bohn, and A.E. Melchinger. 2001. Predicting progeny means and variances of winter wheat crosses from phenotypic values of their parents. *Crop Sci.* 41(5):1470–1478. doi:10.2135/cropsci2001.4151470x
- VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91(11):4414–4423. doi:10.3168/jds.2007-0980
- Witcombe, J.R., S. Gyawali, M. Subedi, D.S. Virk, and K.D. Joshi. 2013. Plant breeding can be made more efficient by having fewer, better crosses. *BMC Plant Biol.* 13(1):22. doi:10.1186/1471-2229-13-22
- Zhao, Y., J. Zeng, R. Fernando, and J.C. Reif. 2013. Genomic prediction of hybrid wheat performance. *Crop Sci.* 53(3):802–810. doi:10.2135/cropsci2012.08.0463
- Zhong, S., and J.L. Jannink. 2007. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics* 177(1):567–576. doi:10.1534/genetics.107.075358

6. DISCUSIÓN GENERAL

Para la aplicación de SG como herramienta de predicción del desempeño de individuos que no han sido fenotipados, es necesario entrenar modelos de SG con una población de entrenamiento conocida y de gran tamaño, que reúna toda la información disponible del programa de mejoramiento. Esto implica utilizar información de varios caracteres, años y ensayos. La información de ensayos multi-ambiente (MET) potencialmente presenta IGA. Por lo tanto, para predecir el comportamiento de nuevos individuos es necesario evaluar la presencia de IGA y estudiar la mejor estrategia para realizar las predicciones. Los resultados presentados en el capítulo 3 de este trabajo muestran que los mega-ambientes se formaron básicamente dividiendo años con bajo rendimiento de años con alto rendimiento. Esta misma división fue propuesta por Windhausen et al. (2012) al analizar la base de datos de maíz de CIMMYT y subdividir los datos según distintos criterios agronómicos. Los autores proponen que la presencia de IGA con cambio de ranking, ocurre únicamente al sub-dividir los ambientes con bajos y altos rendimientos, estableciendo éste como criterio futuro para subdividir los ambientes objetivos.

En este trabajo encontramos que la habilidad predictiva del modelo que incluye la IGA ($GBLUP_{gxe}$) fue mayor que la del modelo que utiliza medias en el conjunto de ambientes ($GBLUP_M$). Lo anterior ocurre cuando se utiliza un tamaño grande ($N \geq 500$) de población de entrenamiento y cuando la IGA es significativa (ej. predicciones entre los distintos años en una localidad). En cambio, cuando la IGA fue menor (ej. predicciones entre localidades en un mismo año), no existieron diferencias entre los modelos (Lado et al. 2016). Esto puede ser explicado porque, el modelo que utiliza las medias asume una correlación de uno entre los ambientes, por lo que, cuando los ambientes están muy correlacionados entre sí, es similar predecir con uno u otro modelo. Burgueño et al. (2012) utilizaron ambientes altamente correlacionados y al igual que lo observado en este trabajo, no obtuvieron diferencias en la habilidad predictiva de los modelos que incorporan o no la IGA. Similares resultados fueron encontrados por Dawson et al. (2013) utilizando datos experimentales, debido a que dicha base de datos estaba compuesta por genotipos que fueron mejorados para un amplio conjunto de ambientes, con probable ausencia de IGA entre sus mega-

ambientes. Sin embargo, al usar datos simulados con distintas correlaciones entre ambientes, Dawson et al. (2013) encontraron ventajas en el uso de modelos que consideran la IGA, cuando la correlación entre ambientes es baja. A su vez, Lopez-Cruz et al. (2015), utilizando datos en los cuales se simuló la presencia de IGA para ambientes contrastantes, obtuvieron mejor habilidad predictiva de los modelos que incorporan la IGA. Lo anterior permite concluir que los modelos que incluyen IGA son mejores en la predicción de nuevos individuos en condiciones de cambio de *ranking* entre los genotipos en presencia de IGA, es decir, cuando la correlación entre ambientes es baja.

La información fenotípica disponible en un programa de mejoramiento es de varios caracteres por lo que, integrar dicha información en la predicción de nuevos individuos en SG podría ser ventajosa para predecir caracteres correlacionados entre sí. Sin embargo, Calus y Veerkamp (2011); Jia y Jannink (2012), pioneros en el uso de modelos multi-carácter en SG, sugieren que estos modelos sólo son mejores para predecir caracteres de baja heredabilidad si se utilizan caracteres correlacionados que presenten alta heredabilidad. Estos resultados son evidentes para predicciones realizadas utilizando datos simulados, pero la ventaja de los modelos multi-carácter fue menor cuando la predicción fue realizada a partir de datos experimentales. En este trabajo, el uso de estos modelos multi-carácter no presentó ventajas en la predicción de nuevos individuos, lo que era esperable dado que los caracteres de calidad utilizados presentan valores de heredabilidad similares.

Los modelos de SG, además de ser utilizados para la predicción del comportamiento de nuevos individuos, pueden aplicarse para predecir el desempeño de individuos (que ya cuentan con cierta información fenotípica) para nuevos caracteres o nuevos ambientes. Esto es posible debido a que los modelos mixtos usados en SG permiten “prestar” información entre líneas, caracteres y ambientes, valiéndose de la matriz de varianza-covarianza que modela las relaciones existentes. Con este fin, se han desarrollado estrategias que utilizan los modelos multi-ambiente y multi-carácter para optimizar los recursos invertidos en un programa de mejoramiento (Rutkoski et al. 2012; Endelman et al. 2014; Guo et al. 2014; Rutkoski et al. 2016).

Por un lado, el desarrollo de modelos multi-ambiente permite predecir el desempeño de líneas en ambientes en los cuáles no han sido evaluadas. Esto es útil para evaluar un mayor número de líneas utilizando los mismos recursos, ya que no es necesario evaluar todas las líneas en todos los ambientes. Endelman et al. (2014) demostraron que en poblaciones bi-parentales es más efectivo evaluar un mayor número de genotipos de forma desbalanceada y luego predecir el comportamiento utilizando información de relacionamiento genético, que invertir más recursos en fenotipar detalladamente menos individuos. A su vez, Burgueño et al (2012) y Lopez-Cruz et al. (2015) encontraron que la predicción de individuos conocidos en nuevos ambientes es más precisa cuando se utilizan modelos que incluyan la IGA, en ambientes correlacionados entre sí.

Los resultados obtenidos muestran que para predecir individuos conocidos en nuevos ambientes es preferible utilizar información de grupos de ambientes con menor IGA, como por ejemplo, mega-ambientes definidos a partir de los datos (Lado et al. 2016). A diferencia de Burgueño et al. (2012), no se obtuvo un modelo que siempre predijera mejor individuos conocidos en nuevos ambientes; en ocasiones fue mejor el modelo $GBLUP_{gxe}$ y en otras el modelo $GBLUP_M$, sin mostrar un patrón claro de comportamiento. En relación con esto, Lopez-Cruz et al. (2015) mostraron que la mejor habilidad predictiva de los modelos que incluyen IGA no es tan clara cuando el conjunto de ambientes presenta una estructura compleja de varianzas-covarianzas, y que en estos casos, es mejor predecir entre pares de ambientes correlacionados. En nuestro caso, la estructura de varianzas-covarianzas dentro de los grupos de ambientes (años, localidades y mega-ambientes) es compleja, ya que existe un amplio rango de correlaciones. Predecir nuevos ambientes utilizando todos los ambientes del grupo, probablemente no sea una buena estrategia, como señalaron Lopez-Cruz et al. (2015). Quizá sea más adecuado seleccionar únicamente los ambientes que muestren altas correlaciones entre sí para predecir genotipos con información fenotípica.

Por otro lado, el uso de modelos multi-carácter es también una herramienta útil en la optimización de recursos de un programa de mejoramiento. Trabajos previos muestran la utilidad de estos modelos para predecir caracteres complejos de medir a través de caracteres correlacionados de fácil medición (Rutkoski et al. 2011; Guo et al.

2014). Los resultados presentados en este trabajo sugieren que la optimización de recursos puede maximizarse fenotipando: 1) una pequeña proporción de las líneas de la población de entrenamiento (30%) para el carácter complejo de medir, y 2) utilizando la población de entrenamiento y la de predicción de forma desbalanceada, para los caracteres correlacionados de fácil medición (Lado et al. 2018).

Lo anteriormente expuesto en conjunto con los resultados publicados en Burgueño et al. (2012) y Endelman et al. (2014), sugieren que puede lograrse una máxima optimización a través del fenotipado desbalanceado de las líneas en distintos ambientes y para distintos caracteres. Dicha estrategia requiere estimar con adecuada precisión las correlaciones entre ambientes y caracteres, para luego predecir los individuos para los ambientes y caracteres faltantes. Esta estrategia, en conjunto con las metodologías de fenotipado de alta precisión (Rutkoski et al. 2012; Montesinos-López et al. 2016; Hayes et al. 2017; Sun et al. 2017) y el uso de la información de pedigrí (Rutkoski et al. 2012; Montesinos-López et al. 2016), pueden mejorar la utilización de los recursos de un programa de mejora genética.

Otra posible aplicación de la SG radica en orientar la selección de los mejores cruzamientos a realizar en el proceso inicial de mejora genética. La selección de los mejores cruzamientos a realizar no necesariamente implica seleccionar los mejores padres en cuanto a su desempeño individual. Por ejemplo, si consideramos dos individuos con buen desempeño, pero genéticamente idénticos, su progenie será exactamente igual a ellos y no habrá lugar para la mejora genética. Por lo tanto, la varianza genética entre los padres incidirá en la varianza de la progenie y por ende, en la selección de los mejores cruzamientos (Souza y Sorrells 1991). Al igual que lo observado por Zhong y Jannink (2007), en este trabajo identificamos que cuando el ratio entre la varianza de las medias de los padres de los posibles cruzamientos y la varianza de los desvíos estándares de las progenies es grande, el mayor determinante de la selección es la media de los padres, o sea el desempeño de cada padre (Lado et al. 2017). En nuestro caso los caracteres de calidad que presentaron valores bajos de este ratio (2,5 – 8,5) mostraron diferencias entre la selección de cruza por la media de los padres y la selección por el 10% superior de la progenie considerando la varianza. En cambio, para la variable rendimiento, que tuvo el mayor valor de este ratio (312),

no se observaron diferencias entre ambas formas de seleccionar los cruzamientos. El rendimiento es una variable compleja influenciada por varios genes, esto explicaría la existencia de mayor variación alélica entre padres, lo cual permitiría obtener mejoras en la siguiente generación incluso sin tomar en cuenta la varianza. Resulta interesante plantear que consecuencias existirían al seleccionar en varias generaciones con una y otra metodología.

Para predecir la varianza de la progenie y seleccionar el 10% superior se consideraron los marcadores en equilibrio o desequilibrio de ligamiento, obteniéndose resultados similares en la selección de cruzamientos (Lado et al. 2017). Sin embargo, es importante considerar que al predecir la varianza considerando el desequilibrio de ligamiento, los puntos de recombinación fueron elegidos al azar sin tomar en cuenta la frecuencia de recombinación entre los marcadores en la población en estudio, consideración que, podría mejorar la estrategia de predicción de la varianza. Se plantea entonces, que evaluar la varianza predicha tomando en cuenta una amplia cobertura de marcadores en el genoma y su desequilibrio de ligamiento podría ser una mejor estrategia para predecir la varianza de la progenie.

Los resultados hallados en este trabajo aportan insumos para la discusión de la estrategia para la inclusión de la SG en el Programa de mejoramiento genético de trigo (PMGT) de INIA.

Este trabajo muestra que en el caso de Uruguay y para la variable rendimiento de grano de trigo, la IGA presente en las localidades no es representativa de la IGA observada entre años, lo que había sido también observado por Ceretta y van Eeuwijk (2008). Las precisiones en las predicciones entre las localidades de un mismo año fueron relativamente altas (precisión media de 0,43 entre todas las localidades), mientras que las precisiones en las predicciones entre años fueron bajas (precisión media de 0,16 entre todos los años). Uno de los probables factores que explicaría la existencia de interacción entre años, es la presencia de fusariosis de la espiga en los años donde el rendimiento fue más bajo (2012 y 2014), debido probablemente a que los mejores genotipos (en rendimiento de grano) fueron genotipos resistentes y *no* genotipos que presentan alelos favorables para los componentes de rendimiento *per se*, como sí ocurrió en los otros años.

Los resultados del análisis de los datos de calidad permitieron definir dos grupos de caracteres correlacionados. Estos grupos coinciden con los propuestos por Vázquez et al. (2012), donde uno de los grupos está formado por caracteres vinculados con la calidad de la proteína de trigo (volumen de sedimentación (SV), parámetro W del alveograma (W) y el pico máximo del mixograma (MH)), mientras que el otro grupo se asocia con caracteres vinculados a la cantidad de proteína (contenido de proteína en gran (Pt), parámetro L del alveograma (L) y contenido de gluten húmedo (WG)). La separación entre estos grupos no es total, ya que también existe relación entre la calidad y la cantidad de proteína. A su vez, las interacciones y relaciones genéticas entre los caracteres de calidad son complejas y aún no han sido plenamente dilucidadas (Simons et al. 2012; Krystkowiak et al. 2017).

El valor de W del alveograma es un parámetro de calidad muy importante, que es difícil de determinar en el laboratorio, dado que, además de sus elevados costos, requieren muestras grandes y mucho tiempo para su determinación (Vázquez 2009). Este trabajo muestra que es posible predecir W utilizando el SV y la MH fenotipados en las poblaciones de entrenamiento y de predicción. El SV y la MH son caracteres de más fácil determinación que el W, en particular el SV (comm. pers. D. Vázquez). A su vez esta predicción puede realizarse optimizando los recursos del programa y utilizando únicamente el 30% de las líneas de la población de entrenamiento sin perder habilidad predictiva, lo que representa una gran ventaja en comparación a utilizar el 100% de las líneas de la población de entrenamiento fenotipadas para W.

Respecto a la selección de cruzamientos los resultados muestran que seleccionar por la media de los valores genéticos de los padres para rendimiento es una buena estrategia para aplicar en el programa de mejoramiento nacional. En cambio para los caracteres de calidad, es más importante considerar la varianza genética existente entre ambos padres.

7. CONSIDERACIONES FINALES Y PERSPECTIVAS

Este trabajo propone que para la aplicación de SG en el mejoramiento genético de trigo, es importante estudiar inicialmente la IGA para seleccionar los modelos adecuados (que incluyan o no IGA según la correlación existente entre ambientes) en el proceso de predicción del desempeño de nuevas líneas experimentales. En este contexto, este trabajo muestra que no existe un único modelo para realizar esta predicción, sino que en algunos casos es más efectivo el $GBLUP_{gxe}$ y en otros el $GBLUP_M$. El modelo $GBLUP_{gxe}$ es útil para predecir entre ambientes con presencia de IGA con cambio de ranking (bajas correlaciones). Sin embargo en algunos casos puede convenir eliminar ambientes con bajas correlaciones disminuyendo la IGA y prediciendo con el modelo $GBLUP_M$.

La inclusión de modelos multi-carácter en la aplicación de SG, no es útil para la predicción del comportamiento de nuevos individuos (individuos sin información fenotípica). Es recomendable predecir un carácter a la vez, dada la simplicidad de los modelos.

La habilidad predictiva de los modelos se mantiene al fenotipar menos individuos (hasta un 30% del total) para un carácter costoso o de difícil medición, si se cuenta con información complementaria de otros caracteres correlacionados fenotipados en todas las líneas (de las poblaciones de entrenamiento y predicción). Además de reemplazar el fenotipado del carácter de interés, también es posible obtener buenas predicciones fenotipando menos individuos (ej. el 50%) para los caracteres correlacionados. La mejor estrategia al fenotipar menos individuos es hacerlo de forma desbalanceada, fenotipar distintas líneas para distintos caracteres. A futuro, los datos de fenotipado en alta escala podrían usarse como caracteres secundarios correlacionados que ayuden en la predicción de los caracteres complejos de interés. Este tipo de fenotipado permitiría reunir información de muchas líneas, que junto con la información genómica y el pedigrí, sería útil para incrementar la precisión de las predicciones de caracteres costosos o de difícil medición.

La optimización de recursos de un programa también puede hacerse evaluando distintos individuos en diferentes ambientes correlacionados entre sí, para luego predecir los individuos faltantes utilizando los modelos multi-ambiente. A futuro,

podrían diseñarse experimentos que consideren la mejor forma de predecir las líneas faltantes tomando en cuenta esta estrategia. Lo anterior evitaría evaluar todas las líneas en todos los ambientes, como se hace hoy en día en los programas de mejoramiento, con su consecuente uso de recursos.

En el caso de la mejora de la variable rendimiento, la selección de los mejores cruzamientos debería realizarse utilizando la media de los padres, dado que no presenta diferencias con la selección por el 10% superior de la progenie predicha y es de más simple aplicación. En cambio, para las variables que determinan la calidad de trigo, predecir la varianza y seleccionar por el 10% superior de la progenie predicha, constituye una mejor estrategia a aplicar en la selección de los cruzamientos. A futuro, parece interesante evaluar la precisión en la predicción de la varianza de la progenie utilizando información genética de alta densidad. Además, sería interesante comparar la selección realizada considerando la varianza de la progenie o la media de los padres, luego de varias etapas de mejoramiento.

8. BIBLIOGRAFÍA

- Akdemir D, Sanchez JI, Jannink JL. 2015. Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution*, 47(1):38
- Allard R. 1980. *Principios de la mejora genética de las plantas*, 4a. ed. Barcelona: Omega
- Andaya VC, Mackill DJ. 2003. Mapping of QTLs associated with cold tolerance during the vegetative stage in rice. *Journal of Experimental Botany*, 54(392):2579-85
- Asoro FG, Newell M a., Beavis WD, Scott MP, Tinker N a., Jannink J-L. 2013. Genomic, Marker-Assisted, and Pedigree-BLUP Selection Methods for β -Glucan Concentration in Elite Oat. *Crop Science*, 53(5):1894
- Asoro FG, Newell M a., Beavis WD, Scott MP, 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(2):132-144
- Baenziger PS, Depauw RM. 2009. *Wheat Breeding: Procedures and Strategies*. En: *Wheat Science and Trade*. Oxford: Wiley-Blackwell, pp 273-308
- Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J. 2016. Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Science*, 242:23-36
- Battenfield SD, Guzmán C, Gaynor RC, Singh RP, Peña RJ, Dreisigacker S, Fritz AK, Poland JA. 2016. Genomic Selection for Processing and End-Use Quality Traits in the CIMMYT Spring Bread Wheat Breeding Program. *The Plant Genome*, 9(2)
- Bayer PE, Ruperao P, Mason AS, Stiller J, Chan CKK, Hayashi S, Long Y, Meng J, Sutton T, Visendi P, Varshney RK, Batley J, Edwards D. 2015. High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*. *Theoretical and Applied Genetics*, 128(6):1039-1047
- Bernardo R. 2014. Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. *Crop Science*, 54(6):2586-2595
- Bernardo R. 2003. Parental selection, number of breeding populations, and size of each population in inbred development. *Theoretical and Applied Genetics*,

107(7):1252-1256

- Bernardo R .1991. Retrospective Index Weights Used in Multiple Trait Selection in a Maize Breeding Program. *Crop Science*, 31(5):1174
- Bertan I, Carvalho FIF De, Oliveira AAC De. 2007. Parental Selection Strategies in Plant Breeding Programs. *Journal of Crop Science and Biotechnology*, 10(4):211-222
- Bonjean A, Angus WJ, Ginkel M van. 2016. The world wheat book : a history of wheat breeding. Volume 3. France: Lavoisier-Tec & doc
- Bonnett DG, Rebetzke GJ, Spielmeier W. 2005. Strategies for efficient implementation of molecular markers in wheat breeding. *Molecular Breeding*, 15(1):75-85
- Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3):314-31
- Brim CA. 1966. A Modified Pedigree Method of Selection in Soybeans1. *Crop Science*, 6(2):220
- Burgueño J, de los Campos G, Weigel K, Crossa J. 2012. Genomic Prediction of Breeding Values when Modeling Genotype \times Environment Interaction using Pedigree and Dense Molecular Markers. *Crop Science*, 52(2):707-719
- Burgueño J, Crossa J, Cornelius PL, Yang R-C. 2008. Using Factor Analytic Models for Joining Environments and Genotypes without Crossover Genotype \times Environment Interaction. *Crop Science*, 48(4):1-15
- Calus MPL, Veerkamp RF. 2011. Accuracy of multi-trait genomic selection using different methods. *Genetics, Selection, Evolution*, 43(1):26
- Carey G. 2000. *Quantitative Genetics II: Estimation and Testing*. En: Brace-Thompson J. (Eds.) *Human Genetics for the Social Sciences*. California: SAGE Publications, Inc.1-16
- Casassola A, Brammer S. 2013. Parental selection of wheat lines based on phenotypic characterization and genetic diversity. *Crop Breeding and Applied Biotechnology*, (13):49-58
- Castro AJ, Gamba F, German S, Gonzalez S, Hayes PM, Pereyra S, 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
- Ceretta S, van Eeuwijk F. 2008. Grain Yield Variation in Malting Barley Cultivars in Uruguay and Its Consequences for the Design of a Trials Network. *Crop Science*, 48(1):167-180
- Ceron-Rojas JJ, Crossa J, Arief VN, Basford K, Rutkoski J, Jarquín D, Alvarado G, Beyene Y, Semagn K, DeLacy I. 2015. A Genomic Selection Index Applied to Simulated and Real Data. *G3:Genes|Genomes|Genetics*, 5(10):2155-2164
- Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S. 2013. Next-generation phenotyping: Requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. *Theoretical and Applied Genetics*, 126(4):867-887
- Condit R, Hubbell SP. 1991. Abundance and DNA sequence of two-base repeat regions in tropical tree genomes. *Genome*, 34(1):66-71
- Cooper M, Messina C, Podlich D. 2014. Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop & Pasture Science*, 65:311-336
- Crain J, Mondal S, Rutkoski J, Singh RP, Poland J. 2018. Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. *The Plant Genome*, 11(1):170043
- Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK. 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, 22(11):961-975
- Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, 112(1):48-60
- Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E, Jannink J-L, Lopez Cruz M a, Babu R.

2013. Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3: Genes|Genomes|Genetics*, 3(11):1903-26
- Crossa J, Cornelius PL, Vargas M. 2010a. Linear, Bilinear, and Linear-Bilinear Models for Analyzing Genotype× Environment Interaction. *Canadian Journal of Plant Science*, 90(5): 561-574
- Crossa J, Campos GDL, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun H-J. 2010b. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 186(2):713-24
- Crossa J, Fox PN, Pfeiffer WH, Rajaram S, Gauch HG. 1991. AMMI adjustment for statistical analysis of an international wheat yield trial. *Theoretical and Applied Genetics*, 81(1):27-37
- Cuesta-Marcos A, Casas AM, Hayes PM, Gracia MP, Lasa JM, Ciudad F, Codesal P, Molina-Cano JL, Igartua E. 2009. Yield QTL affected by heading date in Mediterranean grown barley. *Plant Breeding*, 128(1):46-53
- Davey JW, Hohenlohe P a, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7):499-510
- Dawson JC, Endelman JB, Heslot N, Crossa J, Poland J, Dreisigacker S, Manès Y, Sorrells ME, Jannink J-L. 2013. The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research*, 154:12-22
- de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327-345
- de los Campos G, Gianola D, Rosa GJM. 2009a. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science*, 87(6):1883-7
- de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM. 2009b. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1):375-85

- Dekkers JCM. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. *Journal of Animal Breeding and Genetics*, 124(1961):331-341
- Dekkers JC. 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *Journal of Animal Science*, 82:E313-E328
- DIEA (Dirección de Estadísticas Agropecuarias). 2017. Producción [En línea]. En: Anuario estadístico agropecuario 2015. Montevideo: MGAP (Ministerio de Ganadería, Agricultura y Pesca). Consultado 5 febrero 2018. Disponible en: <http://www2.mgap.gub.uy/portal/page.aspx?2,diea,diea-anuario-2015,O,es,0>
- Dole J, Weber DF. 2007. Detection of Quantitative Trait Loci Influencing Recombination Using Recombinant Inbred Lines. *Genetics*, 2319(4):2309-2319
- Dos Santos JPR, De Castro Vasconcellos RC, Pires LPM, Balestre M, Von Pinho RG. 2016. Inclusion of dominance effects in the multivariate GBLUP model. *PLOS One*, 11(4):e0152045
- Duvick D. 2005. Genetic progress in yield of United States maize (*Zea mays* L.). *Maydica*, 50:193-202
- Eathington SR, Crosbie TM, Edwards MD, Reiter RS, Bull JK. 2007. Molecular Markers in a Commercial Breeding Program. *Crop Science*, 47(Supplement_3):S-154-S-163
- Ellis H, Spielmeyer W, Gale R, Rebetzke J, Richards A. 2002. «Perfect» markers for the Rht-B1b and Rht-D1b dwarfing genes in wheat. *Theoretical and Applied Genetics*, 105(6-7):1038-1042
- Elshire R, Glaubitz J, Sun Q. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLOS One*, 6(5):1-10
- Endelman JB, Atlin GN, Beyene Y, Semagn K, Zhang X, Sorrells ME, Jannink J-L. 2014. Optimal Design of Preliminary Yield Trials with Genome-Wide Markers. *Crop Science*, 54(1):48-59
- Endelman JB, Jannink J-L. 2012. Shrinkage estimation of the realized relationship matrix. *G3: Genes|Genomes|Genetics*, 2(11):1405-13
- Endelman JB. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal*, 4(3):250-255

- Falconer DS, Mackay TFC. 1996. Introduction to quantitative genetics, 4th edn. New York : Ronald Press Company
- FAO (Food and Agriculture Organization). 2017. [En línea]. En: Estadísticas en la Organización de las Naciones Unidas para la alimentación y la Agricultura. Consultado: 23 de enero 2018. Disponible en: <http://www.fao.org/faostat/en/#compare>
- Federer TW. 1955. Experimental Design: Theory and Application, 1st edn. New York: Macmillan
- Fehr W. 1987. Principles of cultivar development Volume 1 Theory and Technique. Iowa: Macmillan
- Fisher RA. 1925. Statistical methods for research workers, 1st edn. Edinburgh: Oliver and Boyd
- Frank Ildiko E, Friedman JH. 1993. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109-135
- Fu YB. 2015. Understanding crop genetic diversity under modern plant breeding. *Theoretical and Applied Genetics*, 128(11):2131-2142
- Ganal MW, Röder MS. 2007. Microsatellite and SNP markers in wheat breeding. En: *Genomics-Assisted Crop Improvement*. Netherlands: Springer, pp 1-24
- Gianola D, Okut H, Weigel K a, Rosa GJ. 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genetics*, 12(1):87-101
- Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R. 2009. Additive genetic variability and the Bayesian alphabet. *Genetics*, 183(1):347-63
- Gianola D, van Kaam JBCHM. 2008. Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178(4):2289-303
- Gianola D, Fernando RL, Stella A. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, 173(3):1761-1776
- Goddard ME, Hayes BJ. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10(6):381-91
- Grabowski PP, Morris GP, Casler MD, Borevitz JO. 2014. Population genomic

- variation reveals roles of history, adaptation and ploidy in switchgrass. *Molecular Ecology*, 23(16):4059-4073
- Graybosch R a., Peterson CJ. 2010. Genetic Improvement in Winter Wheat Yields in the Great Plains of North America, 1959–2008. *Crop Science*, 50(5):1882-1890
- Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G. 2014. Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genetics*, 15(1):30-36
- Gupta PK, Rustgi S, Mir RR. 2008. Array-based high-throughput DNA markers for crop improvement. *Heredity*, 101(1):5-18
- Gupta PK, Varshney RK. 2000. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica*, 113:163-185
- Gupta PK, Varshney RK, C SP, B R. 1999. Molecular markers and their applications in wheat breeding. *Plant Breeding*, 118:369-390
- Gutiérrez L, Cuesta-Marcos A, Castro AJ, von Zitzewitz J, Schmitt M, Hayes PM. 2011. Association Mapping of Malting Quality Quantitative Trait Loci in Winter Barley: Positive Signals from Small Germplasm Arrays. *The Plant Genome Journal*, 4(3):256
- Guzy-Wrobelska J, Szarejko I. 2003. Molecular and agronomic evaluation of wheat doubled haploid lines obtained through maize pollination and anther culture methods. *Plant Breeding*, 122(4):305-313
- Habier D, Fernando RL, Dekkers JCM. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389-2397
- Harlan H V, Martini ML. 1938. The Effect of Natural Selection in a Mixture of Barley Varieties. *Journal of Agricultural Research*, 57(3):189-201
- Hawkesford MJ, Araus J-L, Park R, Calderini D, Miralles D, Shen T, Zhang J, Parry M a. J. 2013. Prospects of doubling global wheat yields. *Food and Energy Security*, 2(1):34-48
- Hayes BJ, Walker JPC, Kant ALCS. 2017. Accelerating wheat breeding for end - use quality with multi - trait genomic predictions incorporating near infrared and nuclear magnetic resonance - derived phenotypes. *Theoretical and Applied Genetics*, 130(12):2505-2519

- Hayes BJ, Cogan NOI, Pembleton LW, Goddard ME, Wang J, Spangenberg GC, Forster JW. 2013. Prospects for genomic selection in forage plant species. *Plant Breeding*, 132(2):133-143
- Hayes BJ, Visscher PM, Goddard ME. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research*, 91(1):47-60
- Hazel LN. 1943. The genetic basis for constructing selection indexes. *Genetics*, 28:476-490
- Heffner EL, Jannink J, Sorrells ME. 2011. Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. *The Plant Genome*, 4(1):65-75
- Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME. 2010. Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. *Crop Science*, 50(5):1681-1690
- Heffner EL, Sorrells ME, Jannink J-L. 2009. Genomic Selection for Crop Improvement. *Crop Science*, 49(1):1-12
- Henderson CR. 1984. *Applications of Linear Models in Animal Breeding Models*. Guelph: University of Guelph. 461
- Henderson CR. 1975. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, 31(2):423
- Henry Y, Buyser J de. 1990. Wheat Anther Culture: Agronomic Performance of Doubled Haploid Lines and the Release of a New Variety "Florin". *Biotechnology in Agriculture and Forestry*, 13:285-352
- Hesey PW, Lantican MA, Dubin HJ, Morris ML. 2002. *Impacts of International Wheat Breeding Research in the Developing Countries, 1966-97*. Mexico D.F.: CIMMYT
- Heslot N, Jannink J-L, Sorrells ME. 2015. Perspectives for Genomic Selection Applications and Research in Plants. *Crop Science*, 55(february):1-12
- Heslot N, Akdemir D, Sorrells ME, Jannink J-L. 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics*, 127(2):463-80
- Heslot N, Sorrells ME, Heslot N, Yang H, Sorrells ME, Jannink J. 2012. Genomic

- Selection in Plant Breeding: A Comparison of Models Genomic Selection in Plant Breeding : *Crop Science*, 52:146-160
- Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, Grondona M, Zambelli A, Windhausen VS, Mathews K, Gorjanc G. 2014. Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. *Crop Science*, 54(4):1476
- Hoerl AE, Kennard RW. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55-67
- Hospital F, Moreau L, Lacoudre F, Charcosset A, Gallais A. 1997. More on the efficiency of marker-assisted selection. *Theoretical and Applied Genetics*, 95(8):1181-1189
- Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T, Dong G, Sang T, Han B. 2009. High-throughput genotyping by whole-genome resequencing. *Genome Research*, 19(6):1068-1076
- Hung H-Y, Browne C, Guill K, Coles N, Eller M, Garcia A, Lepak N, Melia-Hancock S, Oropeza-Rosas M, Salvo S, Upadyayula N, Buckler ES, Flint-Garcia S, McMullen MD, Rocheford TR, Holland JB. 2012. The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity*, 108(5):490-499
- Hyten DL, Cannon SB, Song Q, Weeks N, Fickus EW, Shoemaker RC, Specht JE, Farmer AD, May GD, Cregan PB. 2010. 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
- Ingvarsson PK, Street NR. 2011. Association genetics of complex traits in plants. *The New Phytologist*, 189(4):909-922
- Isidro J, Jannink J-L, Akdemir D, Poland J, Heslot N, Sorrells ME. 2015. Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics*, 128:145-158
- Iwata H, 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

- Jannink J-L, Jansen R. 2001. Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics*, 157(1):445-454
- Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, Burgueño J, de los Campos G. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, 127(3):595-607
- Jia Y, Jannink J-L. 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, 192(4):1513-1522
- Jiang G-L. 2013. Molecular Markers and Marker-Assisted Breeding in Plants. En: *Plant Breeding from Laboratories to Fields*. En: *Plant Breeding from Laboratories to Fields*. Denmark:IntechOpen, pp 45-83
- Kearsey MJ, Farquhar AGL. 1998. QTL analysis in plants ; where are we now ? *Heredity*, 80(2):137-142
- Kebschull JM, Zador AM. 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, 43(21):1-15
- Kirigwi FM, Van Ginkel M, Brown-Guedira G, Gill BS, Paulsen GM, Fritz a. K. 2007. Markers associated with a QTL for grain yield in wheat under drought. *Molecular Breeding*, 20(4):401-413
- Kojima T, Nagaoka T, Noda K, Ogihara Y. 1998. Genetic linkage map of ISSR and RAPD markers in Einkorn wheat in relation to that of RFLP markers. *Theoretical and Applied Genetics*, 96(1):37-45
- Korzun V, Röder MS, Ganai MW, Worland AJ, Law CN. 1998. Genetic analysis of the dwarfing gene (Rht8) in wheat. Part I. Molecular mapping of Rht8 on the short arm of chromosome 2D of bread wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics*, 96(8):1104-1109
- Korzun V, Roder M, Worland AJ, Borner A. 1997. Intrachromosomal mapping of genes for dwarfing (Rht12) and vernalization response (Vrn1) in wheat by using RFLP and microsatellite markers. *Plant Breeding*, 116(3):227-232
- Krystkowiak K, Langner M, Adamski T, Salmanowicz BP, Kaczmarek Z, Krajewski P, Surma M. 2017. Interactions between Glu-1 and Glu-3 loci and associations of selected molecular markers with quality traits in winter wheat (*Triticum aestivum*

- L.) DH lines. *Journal of Applied Genetics*, 58(1):37-48
- Kujur A, Upadhyaya HD, Shree T, Bajaj D, Das S, Saxena MS, Badoni S, Kumar V, Tripathi S, Gowda CLL, Sharma S, Singh S, Tyagi AK, Parida SK. 2015. Ultra-high density intra-specific genetic linkage maps accelerate identification of functionally relevant molecular tags governing important agronomic traits in chickpea. *Scientific Reports*, 5(1):9468
- Kumar S, Banks TW, Cloutier S. 2012a. SNP discovery through next-generation sequencing and its applications. *International Journal of Plant Genomics*, 2012:1-15
- Kumar S, Chagné D, Bink MC a M, Volz RK, Whitworth C, Carlisle C. 2012b. Genomic selection for fruit quality traits in apple (*Malus domestica* Borkh.). *PLOS One*, 7(5):e36674
- Lado B, Vázquez D, Quincke M, Silva P, Aguilar I, Gutiérrez L. 2018. Resource Allocation Optimization with Multi-trait Genomic Prediction for Bread Wheat (*Triticum aestivum* L.) Baking Quality. *Theoretical and Applied Genetics* (Aceptado).
- Lado B, Battenfield S, Guzmán C, Quincke M, Singh RP, Dreisigacker S, Peña RJ, Fritz A, Silva P, Poland J, Gutiérrez L. 2017. Strategies for Selecting Crosses Using Genomic Prediction in Two Wheat Breeding Programs. *The Plant Genome* 10(2):1 - 12
- Lado B, González Barrios P, Silva P, Quincke M, Gutiérrez L. 2016. Modeling Genotype x Environment Interaction for Genomic Selection with Unbalanced Data from a Wheat Breeding Program. *Crop Science*, 56:2165 - 2179
- Lado B, Matus I, Rodríguez A, Inostroza L, Poland J, Belzile F, del Pozo A, Quincke M, Castro M, von Zitzewitz J. 2013. Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3: Genes|Genomes|Genetics*, 3(12): 2105-2114
- Lande R, Thompson R. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3):743-756
- Laurie DA, O'Donoghue LS. 1994. Wheat × Maize Crosses for the Production of Wheat Haploids. En: *Biotechnology in Agriculture and Forestry* 13. Berlin:

Springer. 102-118

- Laurie DA, Bennett MD. 1988. The production of haploid wheat plants from wheat x maize crosses. *Theoretical and Applied Genetics*, 76(3): 393-397
- Lee M. 1995. DNA markers and plant breeding programs. *Advances in Agronomy*, 55:265-344
- Locatelli A, Cuesta-Marcos A, Gutiérrez L, Hayes PM, Smith KP, Castro AJ. 2012. Genome-wide association mapping of agronomic traits in relevant barley germplasm in Uruguay. *Molecular Breeding*, 31(3):631-654
- Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink J, Singh RP, Autrique E, Campos GDL. 2015. Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker x Environment Interaction Genomic Selection Model. *G3: Genes|Genomes|Genetics*, 5(April):569-582
- Lorenz AJ. 2013. Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. *G3: Genes|Genomes|Genetics*, 3(3):481-91
- Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink JL. 2011. Genomic Selection in Plant Breeding. Knowledge and Prospects. *Advances in Agronomy*, 110(C):77-123
- Lorenzana RE, Bernardo R. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics*, 120(1):151-61
- Love HH. 1927. A Program for Selecting and Testing Small Grains in Successive Generations Following Hybridization1. *Agronomy Journal*, 19(8):705
- Lyra DH, Mendonça LDF, Galli G, Alves FC. 2017. Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. *Molecular Breeding*, 37:80-94
- Mackay TFC, Stone E a, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10(8):565-77
- Malosetti M, Ribaut J-M, van Eeuwijk F a. 2013. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology*, 4(44):1-17

- Malosetti M, Ribaut JM, Vargas M, Crossa J, Eeuwijk F a.. 2007. A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L.). *Euphytica*, 161(1-2):241-257
- Malosetti M, Voltas J, Romagosa I, Ullrich SE, Eeuwijk F a Van. 2004. Mixed models including environmental covariables for studying QTL by environment interaction. *Euphytica*, 137(1):139-145
- Manolio T, Collins F, Cox N. 2009. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747-753
- McCouch S. 2004. Diversifying selection in plant breeding. *PLOS Biology*, 2(10):e347
- Melchinger AE, Gumber RK, Leipert RB, Vuylsteke M, Kuiper M. 1998. Prediction of testcross means and variances among F3 progenies of F1 crosses from testcross means and genetic distances of their parents in maize. *Theoretical and Applied Genetics*, 96(3-4):503-512
- Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819-1829
- Mezghani N, Ruess H, Tarchoun N, Ben Amor J, Simon PW, Spooner DM. 2018. Genotyping-by-sequencing reveals the origin of the Tunisian relatives of cultivated carrot (*Daucus carota*). *Genetic Resources and Crop Evolution*, 65(5):1359-1368
- Michel S, Ametz C, Gungor H, Epure D, Grausgruber H, Lschenberger F, Buerstmayr H. 2016. Genomic selection across multiple breeding cycles in applied bread wheat breeding. *Theoretical and Applied Genetics*, 129(6): 1179-1189
- Mohammadi M, Tiede T, Smith KP. 2015. PopVar: A Genome-Wide Procedure for Predicting Genetic Variance and Correlated Response in Biparental Breeding Populations. *Crop Science*, 55(5):2068
- Montesinos-López OA, Montesinos-López A, Crossa J, Toledo FH, Pérez-hernández O, Eskridge KM, Rutkoski J. 2016. A Genomic Bayesian Multi-trait and Multi-environment Model. *G3:Genes|Genomes|Genetics*, 6:2725-2744
- Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES,

- Kresovich S. 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences*, 110(2):453-458
- Nakaya A, Isobe SN. 2012. Will genomic selection be a practical method for plant breeding? *Annals of botany*, 110(6):1303-1316
- Ordas B, Malvar R a, Santiago R, Butron A. 2010. QTL mapping for Mediterranean corn borer resistance in European flint germplasm using recombinant inbred lines. *BMC Genomics*, 11:174-184
- Ortiz R, Trethowan R, Ferrara GO, Iwanaga M, Dodds JH, Crouch JH, Crossa J, Braun HJ. 2007. High yield potential, shuttle breeding, genetic diversity, and a new international wheat improvement strategy. *Euphytica*, 157(3):365-384
- Parisseaux B, Bernardo R. 2004. In silico mapping of quantitative trait loci in maize. *Theoretical and Applied Genetics*, 109(3):508-514
- Paterson A, DeVerna J, Lanini B, Tanksley S. 1990. Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato. *Genetics*, 124:735-742
- Pešek J, Baker RJ. 1969. Desired improvement in relation to selection indices. *Canadian Journal of Plant Science*, 49: 803-804.
- Piepho HP. 2009. Ridge regression and extensions for genomewide selection in maize. *Crop Science*, 49(4):1165-1176
- Piepho H-P. 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics*, 97(1-2):195-201
- Piepho H-P. 1997. Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics*, 53(2):761-766
5(3):103-113
- Poland JA, Rife TW. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. *The Plant Genome Journal*, 5(3):92-102
- Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink J-L. 2012. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome Journal*,

- Prasanna BM, Araus JL, Crossa J, Cairns JE. 2013. High-Throughput and Precision Phenotyping for Cereal Breeding Programs High-Throughput and Precision Phenotyping for Cereal Breeding Programs. En: *Cereal Genomics II*. 341-374
- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945-59
- R Development Core Team. 2016. R: The R Project for Statistical Computing.[En línea]. 23 de enero 2018. <https://www.r-project.org/>
- Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, He Z. 2017. Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. *Molecular Plant*, 10(8):1047-1064
- Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MD V, Kirst M. 2011. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytologist*, 193(3):617-624
- Ribaut J-M, Ragot M. 2007. Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. *Journal of Experimental Botany*, 58(2):351-360
- Rincint R, Oury EKHMFX, Rousset M, Allard V. 2017. Optimization of multi - environment trials for genomic selection based on crop models. *Theoretical and Applied Genetics*, 130(8):1735-1752
- Rincint R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodríguez VM, Moreno-Gonzalez J, Melchinger A, Bauer E, Schoen C-C, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, Moreau L. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics*, 192(2):715-728
- Rutkoski J, Poland J, Mondal S, Autrique E, Pérez LG, Crossa J, Reynolds M, Singh R. 2016. Canopy Temperature and Vegetation Indices from High-Throughput Phenotyping Improve Accuracy of Pedigree and Genomic Selection for Grain Yield in Wheat. *G3:Genes|Genomes|Genetics*, 6(9):2799-2808
- Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL, Sorrells ME.

2015. Efficient Use of Historical Data for Genomic Selection: A Case Study of Stem Rust Resistance in Wheat. *The Plant Genome*, 8(1):1-10
- Rutkoski J, Benson J, Jia Y, Brown-guedira G, Jannink J, Sorrells M. 2012. Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat. *The Plant Genome*, 5(2):51-61
- Rutkoski JE, Heffner EL, Sorrells ME. 2011. Genomic selection for durable stem rust resistance in wheat. *Euphytica*, 179(1):161-173
- Schulthess AW, Yu W, Miedaner T, Wilde P, Reif JC, Zhao Y. 2016. Multiple-trait- and selection indices- genomic predictions for grain yield and protein content in rye for feeding purposes. *Theoretical and Applied Genetics*, 129(2):273-287
- Silva P, Calvo-Salazar V, Condón F, Quincke M, Pritsch C, Gutiérrez L, Castro A, Herrera-Foessel S, von Zitzewitz J, Germán S. 2015. Effects and interactions of genes Lr34, Lr68 and Sr2 on wheat leaf rust adult plant resistance in Uruguay. *Euphytica*, 204(3):599-608
- Simons K, Anderson JA, Mergoum M, Faris JD, Klindworth DL, Xu SS, Sneller C, Ohm JB, Hareland GA, Edwards MC, Chao S. 2012. Genetic mapping analysis of bread-making quality traits in spring wheat. *Crop Science*, 52(5):2182-2197
- Smith AB, Ganesalingam A, Kuchel H, Cullis BR. 2015. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theoretical and Applied Genetics*, 128(1):55-72
- Smith HF. 1936. A Discriminant Function for Plant Selection. *Annual of Eugenics*, 7:240-250
- Snape JW, Riggs TJ. 1975. Genetical Consequences of Single Seed Descent in the Breeding of Self-Pollinating Crops. *Heredity*, 35(2):211-219
- Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE. 2008. Genomic selection using different marker types and densities. *American Society of Animal Science*, 86:2447-2454
- Souza E, Sorrells ME. 1991. Prediction of progeny variation in oat from parental genetic relationships. *Theoretical and Applied Genetics*, 82(2):233-241
- Spindel J, Begun H, Akdemir D, Collard B, Redoña E, Jannink J-L, McCouch S. 2016. Genome - wide prediction models that incorporate de novo GWAS are a powerful

- new tool for tropical rice improvement. *Heredity*, 116:395-408
- Spitters CJT. 1979. Competition and its consequences for selection in barley breeding. En: *Agricultural Research Report*. pp 167-176
- Stuber CW, Lincoln SE, Wolff DW, Helentjaris T, Lander ES. 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
- Subandi WA, Compton A, Empig LT. 1973. Comparison of the efficiencies of selection indices for three traits in two variety crosses of corn. *Crop Science*, 13: 184-186
- Sun J, Rutkoski JE, Poland JA, Crossa J, Jannink J, Sorrells ME. 2017. Multitrait , Random Regression, or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. *The Plant Genome*, 10(2):1-12
- Tanksley SD. 1993. Mapping Polygenes. *Annual Review of Genetics*, 27(1):205-233
- Tanksley SD, Young ND, Paterson AH, Bonierbale MW. 1989. RFLP mapping in plant breeding: New tools for an old science. *Nature Biotechnology*, 7(3):257-264
- Thomson MJ. 2014. High-Throughput SNP Genotyping to Accelerate Crop Improvement. *Plant Breeding and Biotechnology*, 2(3):195-212
- Tibshirani R. 1996. Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B*, 58(1):267-288
- Tiede T, Kumar L, Mohammadi M, Smith KP. 2015. Predicting genetic variance in biparental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. *Molecular Breeding*, 35(10):1-13
- Trethowan RM, Reynolds MP, Ortiz-Monasterio JI, Ortiz R. 2007. The Genetic Basis of the Green Revolution in Wheat Production. En: *Plant Breeding Reviews*. Hoboken, New Jersey: John Wiley & Sons, Inc. 39-58
- Trethowan RM, Reynolds M, Sayre K, Ortiz-Monasterio I. 2005. Adapting wheat cultivars to resource conserving farming practices and human nutritional needs. *Annals of Applied Biology*, 146(4):405-413

- Uitdewilligen JGAML, Wolters AMA, D'hoop BB, Borm TJA, Visser RGF, van Eck HJ. 2013. A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. *PLOS One*, 8(5):e62355
- Utz HF, Bohn M, Melchinger AE. 2001. Predicting progeny means and variances of winter wheat crosses from phenotypic values of their parents. *Crop Science*, 41(5):1470-1478
- VanRaden PM. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414-4423
- Varshney RK, Tuberosa R. 2007. *Genomics-Assisted Crop Improvement. Vol 2: Genomics Applications in Crops*. Netherlands: Springer
- Vázquez D, Berger AG, Cuniberti M, Bainotti C, Zavariz de Miranda M, Scheeren PL, Jobet C, Zúñiga J, Cabrera G, Verges R, Peña RJ. 2012. Influence of cultivar and environment on quality of Latin American wheats. *Journal of Cereal Science*, 56(May 2016):196-203
- Vázquez D. 2009. Aptitud industrial de trigo. In: *Inst. Nac. Investig. Agropecu.* [En línea]. 23 de enero 2018. Disponible en: [http://www.inia.uy/Publicaciones/Documentos compartidos/18429130709133540.pdf](http://www.inia.uy/Publicaciones/Documentos_compartidos/18429130709133540.pdf)
- Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era--concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255-66
- von Zitzewitz J, Cuesta-Marcos A, Condon F, Castro AJ, Chao S, Corey A, Filichkin T, Fisk SP, Gutierrez L, Haggard K, Karsai I, Muehlbauer GJ, Smith KP, Veisz O, Hayes* PM. 2011. The Genetics of Winterhardiness in Barley: Perspectives from Genome-Wide Association Mapping. *The Plant Genome Journal*, 4(1):76-91
- Vuylsteke M, Peleman JD, van Eijk MJ. 2007. AFLP technology for DNA fingerprinting. *Nature Protocols*, 2(6):1387-1398
- Wang X, Li L, Yang Z, Zheng X, Yu S, Xu C, Hu Z. 2017. Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Heredity*, 118(3):302-310

- Wang DGD., Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES, DG W, JB F, CJ S. 1998. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science*, 280(5366):1077-1082
- Watson A, Ghosh S, Williams MJ, Cuddy WS, Simmonds J, Rey MD, Asyraf Md Hatta M, Hinchliffe A, Steed A, Reynolds D, Adamski NM, Breakspear A, Korolev A, Rayner T, Dixon LE, Riaz A, Martin W, Ryan M, Edwards D, Batley J, Raman H, Carter J, Rogers C, Domoney C, Moore G, Harwood W, Nicholson P, Dieters MJ, Delacy IH, Zhou J, Uauy C, Boden SA, Park RF, Wulff BBH, Hickey LT. 2018. Speed breeding is a powerful tool to accelerate crop research and breeding. *Nature Plants*, 4(1):23-29
- White JW, Andrade-Sanchez P, Gore M a., Bronson KF, Coffelt T a., Conley MM, Feldmann K a., French AN, Heun JT, Hunsaker DJ, Jenks M a., Kimball B a., Roth RL, Strand RJ, Thorp KR, Wall GW, Wang G. 2012. Field-based phenomics for plant genetics research. *Field Crops Research*, 133:101-112
- Whittaker JC, THOMPSON R, DENHAM MC. 1999. Marker-assisted selection using ridge regression. *Annals of Human Genetics*, 63(4):366-366
- Williams JS. 1962. The evaluation of a selection index. *Biometrics*, 18: 375-393.
- Windhausen VS, Wagener S, Magorokosho C, Makumbi D, Vivek B, Piepho H-P, Melchinger AE, Atlin GN. 2012. Strategies to Subdivide a Target Population of Environments: Results from the CIMMYT-Led Maize Hybrid Testing Programs in Africa. *Crop Science*, 52(5):2143-2152
- Witcombe JR, Gyawali S, Subedi M, Virk DS, Joshi KD. 2013. Plant breeding can be made more efficient by having fewer, better crosses. *BMC Plant Biology*, 13(1):22-34
- Witcombe J., Hollington P., Howarth C., Reader S, Steele K.. 2008. Breeding for abiotic stresses for sustainable agriculture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1492):703-716
- Xu S. 2003. Estimating polygenic effects using markers of the entire genome.

Genetics, 163(2):789-801

- Yi N, Xu S. 2008. Bayesian LASSO for Quantitative Trait Loci Mapping. *Genetics*, 1055:1045-1055
- Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203-208
- Zäim M, El Hassouni K, Gamba F, Filali-Maltouf A, Belkadi B, Sourour A, Amri A, Nachit M, Taghouti M, Bassi FM. 2017. Wide crosses of durum wheat (*Triticum durum* Desf.) reveal good disease resistance, yield stability, and industrial quality across Mediterranean sites. *Field Crops Research*, 214:219-227
- Zhang X, Sallam A, Gao L, Kantarski T, Poland J, DeHaan LR, Wyse DL, Anderson JA. 2016. Establishment and Optimization of Genomic Selection to Accelerate the Domestication and Improvement of Intermediate Wheatgrass. *The Plant Genome*, 9(1):1-18
- Zhong S, Dekkers JCM, Fernando RL, Jannink J-L. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. *Genetics*, 182(1):355-64
- Zhong S, Jannink JL. 2007. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics*, 177(1):567-576