UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE AGRONOMÍA

MAPEO ASOCIATIVO MULTI-CARÁCTER MULTI-AMBIENTE PARA VARIABLES AGRONÓMICAS Y FISIOLÓGICAS EN TRIGO

por

Sofía Patricia BRANDARIZ ZERBONI

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

MONTEVIDEO
URUGUAY
agosto 2015

Tesis aprobada por el tribunal integrado por el Dr. Marcos Malosetti, el Dr. Antonio Augusto F. García, y la Dra. Virginia Gravina, el 07 de agosto de 2015. Autora: Ing. Agr. Sofía Brandariz. Directora Dra. Lucía Gutiérrez, Co-directores: Dr. Omar Borsani y Dr. Martín Quincke.

AGRADECIMIENTOS

Me gustaría agradecer a todos los que participaron de este proceso en mayor o menor medida e hicieron que fuera una gran instancia de aprendizaje para mí, que si bien no siempre fue fácil, me brindaron el apoyo necesario para continuar en el mismo. En primer lugar a mi tutora, Lucía Gutiérrez, quien me brindó inicialmente la oportunidad de trabajar junto a ella, y que me siguió en todo el proceso, contestando todas mis dudas apresuradas, dándome consejos de todo índole, apoyándome cuando estaba confundida, y aportando el mejor de su conocimiento para poder realizar esta investigación. Asimismo, quisiera agradecer a mis cotutores Omar Borsani y Martín Quincke por aceptar participar de esta tesis. Omar me brindó especial apoyo en la interpretación de los resultados obtenidos y me guió en la escritura de unos de los artículos aquí presentados, siempre con la mejor disposición. Martín me brindó la oportunidad de participar de este proyecto confiando en el equipo de investigación en el cual me encuentro, y aportó un enfoque agronómico de gran relevancia a la hora de aplicar los conceptos evaluados en un programa de mejoramiento. También quisiera agradecer a mi comité de seguimiento, Marcos Malosetti y Antonio Augusto F. García, cuyos aportes fueron de gran valor para el desarrollo de la tesis; y a la integrante del tribunal de la defensa de tesis y directora del Departamento de Biometría, Estadística y Computación Virginia Gravina, que me brindó un gran apoyo a lo largo de este emprendimiento.

A su vez, quiero agradecer al Departamento de Biometría, Estadística y Computación de la Facultad de Agronomía, UdelaR, donde desarrollé la tesis, quienes no sólo me ayudaron a analizar e interpretar los datos, sino que fueron un apoyo importantísimo a lo largo de la elaboración de esta tesis. De este grupo quiero resaltar la ayuda de Bettina Lado, quien no solamente fue otra guía junto con mis tutores, sino que su aporte me ayudó también a formarme como persona. Asimismo, quiero destacar la ayuda de Agustín González quien fue otra figura de gran apoyo en la elaboración de mi tesis, siempre estando ahí para cuando precisaba una mano. Juan

Rosas, Natalia Berberian y Andrea Garay también fueron de gran relevancia en los momentos más complicados de la tesis.

Asimismo, quiero agradecer a mis padres y hermanas, siempre apoyándome fuera cual fuera la desición que tomara, brindándome consejos para darme la mejor guía, y dándome cariño en los mejores y peores momentos. Por último pero no menos importane, quiero agradecer a Guillermo Pérez, por ser mi pilar junto con mi familia, ya que sin él nada de esto hubiera sido posible, gracias por estar a mi lado.

TABLA DE CONTENIDO

Página

RESUMEN

Comprender las bases genéticas de variables asociadas al rendimiento en trigo mediante el mapeo asociativo, puede mejorar la productividad del mismo. Imputar la matriz genotípica del mapeo cuando no se tiene un panel de referencia, puede afectar la calidad de ésta y disminuir la performance del mapeo. Los objetivos de nuestro trabajo fueron: comparar la performance del mapeo al imputar la matriz genotípica cuando no hay un panel de referencia y existe una gran proporción de datos faltantes como en genotipado por secuenciación (GBS); y evaluar los factores genéticos asociados a variables relacionadas al rendimiento en trigo considerando la interacción genotipo por ambiente. Para el objetivo uno, evaluamos un panel de GBS de trigo y encontramos que la matriz utilizada para simular los efectos de los QTL (*Quantitative Trait Loci*) afectaba la performance del mapeo. Adicionalmente, evaluamos una matriz genotípica sin datos faltantes de cebada, generamos datos faltantes y detectamos que la performance del mapeo disminuía cuando se realizaba con matrices imputadas. Evaluando la imputación en datos fenotípicos reales, encontramos que había diferencias entre los métodos, concluyendo que no imputar es la mejor opción para realizar el mapeo. Para el objetivo dos, el mismo panel genotípico de trigo de GBS fue utilizado. Los datos fenotípicos se evaluaron en Santa Rosa-Chile y Cauquenes-Chile en 2011 y 2012, midiéndose dieciséis variables fenotípicas. Se realizó un análisis de mapeo: (1) multi-carácter para grupos de variables (componentes del rendimiento, variables asociadas a la hoja y variables morfológicas y fenológicas), (2) multi-ambiente para algunas de las variables medidas. El análisis de mapeo multi-carácter detectó QTL de igual efecto y dos interacciones QTL por variable, y el análisis de mapeo multi-ambiente detectó QTL de igual efecto. Estos resultados pueden contribuir a una mejor comprensión de las bases genéticas del rendimiento en trigo, aportando una primera base para incorporar nuevos QTL en los programas de mejoramiento involucrados.

Palabras clave: Imputación; GBS; QTL; poder; falsos positivos

MULTI-TRAIT AND MULTI-ENVIRONMENT GWAS ANALYSIS FOR AGRONOMICAL AND PHYSIOLOGICAL TRAITS IN WHEAT

SUMMARY

Understanding the genetic basis of yield-related traits in wheat using Genome-Wide Association mapping (GWAS), allows improving wheat's productivity. Imputing the genotypic marker scores when a reference panel is not available can affect the quality of it decreasing the performance of the GWAS analysis. The objectives of this study were: to compare the performance of GWAS analysis when the genotypic marker scores is imputed without a reference panel and there is a large proportion of missing data like in genotyping by sequencing (GBS); and to understand the genetic factors of yield-related traits including genotype by environment interaction. For objective one, we evaluated a wheat GBS panel and we found that the genotypic marker scores used to simulate the QTL (*Quantitative Trait Loci*) affected the performance of the GWAS analysis. Additionally, we evaluated a complete barley genotypic marker scores, we generated missing data and we detected that the GWAS performance decreased when we used the imputed marker scores. When imputing the genotypic markers scores using a real dataset, we found that there were differences between the methods, concluding that not imputing is the best choice for the GWAS analysis. For objective two, the same wheat GBS panel was used. The phenotypic data was obtained in Santa Rosa-Chile and Cauquenes-Chile in 2011 and 2012, and sixteen phenotypic traits were measured. Multi-trait GWAS analysis was performed for groups of traits (yield components, leaf related traits and morphology and phenology traits), and Multi-environment GWAS analysis for some of the traits measured. Main effect QTL and two QTL by traits interactions were detected for the multi-trait GWAS analysis, and main effect QTL were detected for the multi-environment GWAS analysis. These results can contribute to understand wheat yield's genetic basis, providing a first base to incorporate QTL in the breeding programs involved.

Keywords: Imputation; GBS; QTL; power; false positives

1. INTRODUCCIÓN

El trigo (*Triticum aestivum* L.) es el tercer cultivo más importante en términos de producción total mundial con 670 millones de ton producidas en 2012 (FAOSTAT, 2014). A nivel nacional, es el principal cereal invernal con una producción que varió entre 343 mil ton en 1993 a 1.5 millones de ton en 2013, con un máximo en 2011 de 2 millones de ton (FAOSTAT, 2014).

El trigo es una especie autógama (Martin 1990) y alohexaploide (AABBDD), producto de la hibridación de 2 especies: *Triticum turgidum* (el cual aporta los genomas A y B) y *Aegilops tauschii* (que aporta el genoma D) que presenta menor nivel de diversidad (Chao et al. 2010). Su fórmula genómica es 2n=6x=42, con un tamaño del genoma cinco veces superior al del genoma humano (16Gb) y un 25-30 % de sus genes duplicados (Dubcovsky et al. 1996). Si bien el trigo fue uno de los primeros cultivos en domesticarse y sigue permaneciendo entre los más relevantes para la alimentación humana desde su disipación a nivel mundial (Dubcovsky and Dvorak 2007), la seguridad alimenticia podría verse comprometida por el aumento de la demanda alimenticia debido al crecimiento de la población (Mueller et al. 2012) y por el cambio climático (Ewert et al. 2005). A pesar de que el mejoramiento genético vegetal ha conseguido incrementar el rendimiento en grano de trigo de forma exitosa (Fischer 2007), la tasa de incremento del mismo ha disminuido en las últimas décadas (Acreche et al. 2008, Reynolds et al. 2012, Bustos et al 2013). Por lo tanto, la mejora de la productividad del trigo es clave para responder al aumento de la demanda alimenticia y al cambio climático, pero disminuyendo el impacto en la huella ambiental (Mueller et al. 2012). Dado que la demanda mundial de trigo está creciendo a un ritmo más rápido que las ganancias genéticas obtenidas (Barnabás et al. 2008, García et al. 2013), nuevas estrategias de mejoramiento podrían implementarse (Fischer 2007), siendo la biología molecular y la fisiología del cultivo dos disciplinas candidatas para este objetivo (Slafer y Araus 2005).

El rendimiento en grano (*per se*) es un objetivo desafiante porque es un carácter complejo determinado por varios genes (Slafer y Araus 2005, Alimi et al. 2012), y compuesto por otros caracteres que son: granos por espiga, peso de grano y espigas por superficie (Kjaer y Jensen 1996). El rendimiento es a su vez afectado

por la incidencia y severidad de diferentes enfermedades, especialmente por *Puccinia graminis tritici*, *Puccinia tritici* (Chen 2005, Singh et al. 2008) y *Fusarium graminearum* (Windels 2000, Jansen et al. 2005), que implican pérdidas importantes del mismo. Caracteres agronómicos como el rendimiento son en general sensibles a la interacción genotipo por ambiente, lo que significa que la superioridad de los genotipos es relativa al ambiente en el que se desarrollen (Hayes et al. 1993; Boer et al. 2007; Mathews et al. 2008; van Eeuwijk et al. 2010; Malosetti et al. 2013; Alimi et al. 2013). En resumen, el rendimiento, si bien es el objetivo del mejoramiento genético más importante, es un carácter complejo codificado por muchos genes con influencia ambiental y determinado por muchos factores. En consecuencia, la mejora del rendimiento mediante el análisis de la base genética de los caracteres agronómicos y fisiológicos relacionados al rendimiento en grano, podría proporcionar una mejor comprensión del comportamiento del mismo (Slafer y Araus 2005, Fischer 2007), conduciendo a ganancias genéticas mayores.

1.1. MAPEO ASOCIATIVO

El mapeo asociativo (GWAS) se puede utilizar para analizar la base genética del rendimiento en grano y caracteres fisiológicos correlacionados, mediante la búsqueda en todo el genoma de asociaciones marcador-carácter que puedan deberse al desequilibrio de ligamiento (LD, Zhu et al. 2008). El LD es el grado de asociación no aleatoria entre alelos de distintos loci (Yu y Buckler 2006; Zhu et al. 2008), es decir la proporción de gametos que no segregan al azar. El objetivo del GWAS análisis radica en identificar marcadores de herencia simple próximos a factores que afectan características del tipo cuantitativas (Jannink et al. 2009). Su diferencia con el ligamiento físico consiste en que éste refiere a correlaciones físicas entre loci en un cromosoma, mientras que el LD refiere a correlaciones entre alelos en una población (Flint-Garcia et al. 2003). El análisis de GWAS tiene ciertas ventajas. En principio, estudia la herencia compartida de una colección de individuos sin requerimientos específicos de parentesco, donde se ha dado lugar a una gran recombinación (Yu y Buckler 2006), llevando a que no se deban diseñar cruzamientos específicos y que se pueda mapear en germoplasma que sea relevante

para programas de mejoramiento (Malosetti et al. 2007a). En segundo lugar, permite estudiar genotipos con varias generaciones de recombinación donde los principales mecanismos que provocan el LD son la mutación y deriva, mientras que la recombinación lo reduce (Jannink et al., 2009). En tercera instancia, permite tener mayor diversidad alélica y poseer fenotipado disponible de los programas de mejoramiento para la realización del análisis (Malosetti et al. 2007a). La principal limitante de este tipo de análisis radica en la necesidad de controlar por falsos positivos debidos a la estructura poblacional o relacionamiento familiar (Yu y Buckler 2006). Muchos modelos fueron propuestos para estudiar la asociación marcador-carácter (Gutiérrez et al. 2011). Modelos mixtos que controlan por la estructura poblacional han sido exitosamente utilizados para el análisis de GWAS (Yu et al. 2006, Malosetti et al. 2007a, Gutiérrez et al. 2011). Entre ellos, el de Malosetti et al. (2007a) es el más parsimonioso: $y = X\beta + P\underline{v} + e$, donde y: vector fenotipo, X: matriz de marcadores moleculares (genotipos), β: vector desconocido de efectos alélicos a estimar, P: matriz de componentes principales que corrige por estructura, v: vector de las predicciones de los efectos poligénicos aleatorios a estimar, e: error residual. Asimismo, podría incluirse en este modelo las correlaciones genéticas entre caracteres correlacionados, ya que permitiría detectar QTL pleiotrópicos, a través de un GWAS multi-carácter (Malosetti et al. 2007b).

Varios estudios han puesto en práctica el mapeo de QTL multi-carácter y/o multi-ambiente de una manera exitosa (Boer et al. 2007, Malosetti et al. 2007b, Alimi et al. 2013, Malosetti et al. 2013, El-Soda et al. 2014). El mapeo de QTL multi-carácter radica en modelar las correlaciones genéticas entre los caracteres, resultando en la posible detección de QTL pleiotrópicos o ligados mediante la incorporación de una matriz de varianza-covarianza entre los efectos genéticos aleatorios; ya que los efectos genéticos no serán independientes en caso que los QTL estén ligados o sean pleiotrópicos (Malosetti et al. 2007b). Un concepto similar se aplica al mapeo de QTL multi-ambiente, donde los efectos de los QTL entre ambientes pueden cambiar en dirección y magnitud (Malosetti et al. 2007b), dependiendo la respuesta genotípica del ambiente en que el genotipo se desarrolle (Malosetti et al. 2013).

1.2. TECNOLOGÍAS DE SECUENCIACIÓN CON TÉCNICAS DE NUEVA GENERACIÓN

Para que el análisis de GWAS se realice de forma eficiente, los datos tanto fenotípicos como genotípicos deben ser de excelente calidad. Tradicionalmente, el desarrollo de marcadores moleculares del tipo microsatélites implicaba la dedicación de un gran tiempo y costo de desarrollo que fueron disminuidos por la creación de los chips de SNPs (*Single Nucleotide Polymorphism*), pero que presentan la particularidad de ser específicos a la población en la que fueron creados (Davey et al. 2011). Las tecnologías de secuenciación con técnicas de nueva generación (NGS), superan esta limitante y permiten descubrir, secuenciar y genotipar miles de SNPs de todo el genoma en un solo paso (Davey et al. 2011). Se basan en enzimas de restricción, cuya diversidad (variaciones en longitud, en la sensibilidad a la metilación, entre otros) las convierten en una herramienta versátil (Davey et al. 2011). Los tres pasos básicos para genotipar por NGS consisten en: digestión de múltiples muestras de ADN genómico con enzimas de restricción, selección o reducción de los fragmentos resultantes y secuenciación de los fragmentos seleccionados (Davey et al. 2011). Luego, la bioinformática permite obtener la matriz de SNPs a partir de los datos secuenciados (Elshire et al. 2011). Los SNPs obtenidos por NGS se utilizan en análisis como el de diversidad genética, GWAS y Selección Genómica (GS).

Dentro de las técnicas de NGS, una de ellas consiste en el genotipado por secuenciación (GBS), que utiliza la metodología de secuenciación de baja cobertura para genotipado. Esta metodología implica secuenciar muchos marcadores a baja cobertura por individuo, sabiendo que diferentes conjuntos de marcadores se genotiparán por individuo (Davey et al. 2011). El GBS radica en digerir el ADN con enzimas de restricción, colocarle a los fragmentos resultantes adaptadores con códigos de barra (que identifican muestras) y adaptadores para PCR (*Polymerase Chain Reaction*), reunir todas las muestras, amplificar los fragmentos por PCR y secuenciar a los fragmentos que presentar ambos tipos de adaptadores y no son mayores a 1kb (Davey et al. 2011). Esta técnica se ha utilizado exitosamente para

genomas complejos como el del trigo (Poland et al. 2012a, Lado et al. 2013). Dicha utilidad se debe a la propiedad de la técnica de GBS de evitar las regiones repetitivas del genoma usando enzimas de restricción sensibles a la metilación, por lo que mejora la eficiencia de la secuenciación simplificando los problemas de alineación de genomas con varios niveles de complejidad (Elshire et al. 2011).

1.3. MÉTODOS DE IMPUTACIÓN DE LA MATRIZ GENOTÍPICA

Por ser GBS una técnica de secuenciación de baja cobertura para genotipado, una gran cantidad de datos faltantes se obtiene del uso de la misma (GBS ofrece miles de SNPs, pero la mayoría de ellos con una gran proporción de datos faltantes). Por ello se han utilizado diferentes métodos de imputación para la realización de análisis relacionados al mejoramiento molecular, donde las imputaciones de SNPs son requeridas para la GS y han sido utilizadas de forma exitosa (Poland et al. 2012b, Rutkoski et al. 2013, Lado et al. 2013). Un tipo de método de imputación se basa en un panel de referencia completamente genotipado y en el desequilibrio de ligamiento (LD) entre las líneas del panel de referencia y las de las muestras a evaluar. Brevemente, se sustenta principalmente en la copia de segmentos de haplotipos de un panel de referencia densamente genotipo, en individuos genotipados en un subconjunto de dicha referencia (Browing 2008, Jannink et al. 2009, Howie et al. 2011). La precisión de este tipo de método depende de: el LD, de la frecuencia del alelo menor (MAF), de la mínima distancia al marcador molecular no imputado y del grado de diferenciación de las subpoblaciones (Pei et al. 2008, Iwata y Jannink 2010).

Otro tipo de método de imputación se realiza cuando no se presenta dicho panel de referencia. Dentro del segundo tipo de imputación se encuentran los métodos: multivariado normal de maximización de la esperanza (MVN-EM, (Poland et al. 2012b), que considera la matriz de relacionamiento entre individuos e incorpora un método de maximización de la esperanza para calcular los estimadores de máxima verosimilitud de los parámetros desconocidos a estimar asumiendo que cada genotipo se distribuye normal multivariado; y el de imputación por la media, que es el más simple y se basa en imputar por el alelo más común de la población para cada

marcador. Utilizar un método basado en dos pasos (primero imputar y luego mapear sin considerar el error de imputación), puede introducir error en el mapeo y resultar en menor poder de detección de QTL y mayor tasa de falsos positivos.

1.4. PREGUNTAS DE INVESTIGACIÓN Y OBJETIVOS

Las preguntas de investigación planteadas en esta tesis fueron: (1) imputar la matriz genotípica proveniente de GBS sin panel de referencia en un análisis de GWAS, afectará la performance del análisis de GWAS en términos de poder de detección de QTL y tasa de falsos positivos?; (2) permitirá el análisis de GWAS multi-carácter para caracteres asociados a rendimiento, identificar QTL pleiotrópicos o ligados, mejorando el poder de detección de QTL?; y (3) permitirá el análisis de GWAS multi-ambiente identificar QTL con efectos iguales o contrastantes para los ambientes?

Los objetivos planteados en esta tesis fueron: (1) comparar la performance de la imputación de la matriz genotípica en el análisis de GWAS cuando no hay un panel de referencia y una gran proporción de datos faltantes se presenta como en GBS; (2) evaluar los factores genéticos asociados a variables agronómicas y fisiológicas en trigo considerando la interacción genotipo por ambiente.

2. ASCERTAINMENT BIAS WHEN EVALUATING IMPUTATION METHODS[1]

2.1. ABSTRACT

2.1.1. Background: Whole-genome genotyping techniques like Genotyping-by-sequencing (GBS) are being used for genetic studies such as Genome-Wide Association (GWAS) and Genome-Wide Selection (GS). Since GBS generates large amount of missing data, different strategies for imputation have been developed, especially for situations where complete dataset is required like GS. Nevertheless, imputation error may lead to poor performance (i.e. smaller power or higher false positive rate) when complete data is not needed. The aim of this study was to compare the performance of GWAS analysis for major and minor Quantitative Trait Loci (QTL) using different imputation methods, when no reference panel is available and there is a large proportion of missing data like in GBS panels.

2.1.2. Results: In this study we compared the power and false positive rate of QTL detection for imputed and not-imputed marker scores matrices in the following datasets: (1) a complete molecular marker barley panel array, and (2) a GBS wheat panel with an average of 50% missing data. We found that there is an ascertainment bias in method selection. Simulating over a complete matrix and creating missing data at random proved that imputation methods (i.e. mean imputed and Multivariate Normal Expectation Maximization, MVN-EM) have a poorer performance (i.e. smaller power or higher false positive rate). Additionally, we compared if simulating with the different marker scores matrices and performing GWAS with those matrices detected differences. We found that when QTL were simulated with imputed data, the imputation methods performed better but the not-imputed method performed better when simulating with the not-imputed data. Moreover, higher differences between imputation methods were detected for major QTL, and low detection of minor QTL was found. We also compared the different marker scores matrices for GWAS analysis in a real wheat phenotype data, and we found that differences were

[1] Artículo a publicar en: BMC Genomics

neglected between MVN-EM and not-imputed methods indicating that imputing did not improved the GWAS performance.

2.1.3. Conclusions: Poorer performance was found in GWAS analysis when an imputed marker scores matrix was used, no reference panel is available and a there is large proportion of missing data (50%).

2.2. BACKGROUND

Genetic markers are nowadays an essential part of plant and animal breeding programs. Next-generation sequencing (NGS) techniques allow discovering, sequencing and genotyping thousands of Single Nucleotide Polymorphism (SNPs) covering the whole genome in one step [1]. These SNPs are being used in analyses like genetic diversity analysis [2], GWAS [3] and GS [4]. Genotyping-by-sequencing (GBS) is one of the NGS techniques, developed originally for barley and maize, and extended to other complex genomes species like wheat [2–5]. GBS relies on methylation-sensitive restriction enzymes and is therefore highly efficient [6]. Nevertheless, GBS generates a large proportion of missing data when alleles are created due to the use of shorts reads [6]. Therefore, different strategies to impute missing data have been developed and used for genetic analyses [3]. Some imputation methods use reference panels and are based on Linkage Disequilibrium (LD), while other methods not require reference panels. In the first group the most common methods are known as MACH [7], IMPUTE [8], fastPHASE [9], PLINK [10] and Beagle [11]. All use haplotypes segments from a reference panel densely genotyped to imputed missing markers [12–14]. Briefly, MACH uses a Markov Chain based algorithm to infer pairs of haplotypes for each individual's genotypes [7]. IMPUTE consider the sequence of pairs of known haplotypes as hidden states and then models the sequence of hidden state based on a recombination map estimated from the reference data, and finally predicts unknown genotypes [8]. The fastPHASE algorithm is a haplotype clustering algorithm that samples missing genotypes based on allele frequencies estimated from reference haplotypes, and then an Expectation-Maximization (EM) algorithm is used to estimate parameter values to infer missing genotypes [9]. PLINK predicts missing data by the local haplotypic background and by the haplotype formed by the two or more flanking SNPs [10]. Finally, Beagle is a haplotype clustering based algorithm, which uses the localized haplotype cluster model to cluster haplotypes at each marker and then finds the most likely haplotype pairs based on the individual's known genotypes [11]. Therefore, strong LD among markers and low minor allele frequency (MAF) is required for LD imputation methods [15]. Additionally, large genome coverage to decrease distances

among markers, and small population structure is also desirable to ensure imputation accuracy [16]. The second group of methods includes mean imputation, multivariate-normal expectation-maximization (MV-EM) algorithm and random forest. In the mean imputation, the most common allele at a particular marker in the population is used to impute missing data. MVN-EM considers the realized additive relationship matrix between the lines and an EM approach assuming that marker genotypes follow a multivariate normal distribution designed for use with GBS. Random forest is an algorithm that uses multiple decision trees to determine a prediction value for each missing data point. For an overview see [4].

Several studies found that imputation can improve QTL power detection [17, 18], but other studies found that large power is accompanied by larger false positives or by an increase in the multiple-testing penalty [14, 19]. Unless a 'one-hit' procedure is used (i.e. the uncertainty of genotypic probability distributions due to the imputation is incorporated in the GWAS analysis), large imputation error can be generated [20]. Other studies found that imputation should be carefully evaluated because quality control of the data is an important source of loss of power [21]. To carry on GWAS analysis, where one marker at a time is being tested, marker-trait associations can be estimated without marker imputation.

The aim of this study was to compare the performance of imputation methods of the marker scores matrix for GWAS analysis, when there is no reference panel for the lines and markers evaluated, and there is a large proportion of missing data like in GBS panels. Specifically, our objectives were: (1) to evaluate the effect of imputation using a golden standard (i.e. simulation over a complete marker scores matrix), to determine whether ascertainment bias is responsible for imputation success; (2) to evaluate whether the outcome of the imputation performance is affected by the marker scores matrix used to simulate the QTL; and (3) to compare the effect of imputation in a real phenotype wheat panel using GBS data and four phenotypic traits.

2.3. RESULTS

To evaluate the effect of imputation in GBS data when there is no reference panel, we pursued the following strategies. First, to evaluate the effect of imputation using a golden standard (i.e. complete marker scores matrix), we simulated QTL on top of the complete marker scores matrix to create vectors of phenotypic values with Y_{sim} = *NoNA*. Then, we randomly generated the missing values, imputed with the different methods, pursued the GWAS analysis using the different matrices: GMat = *NImp*, when a not-imputed marker scores matrix was used, GMat = *MVN-EM* when an imputed with MVN-EM method [4] matrix was used, and GMat = *Mean* when an imputed by the mean marker scores matrix was used; and then we evaluated the performance. Additionally, for evaluating a possible ascertainment bias we used the golden standard matrix (Y_{sim} = *NoNA*), randomly generated the missing values, imputed with the different methods and then simulated the QTL on top of the different marker scores matrices to create vectors of phenotypic values (Y_{sim} = *NImp,* Y_{sim} = *MVN-EM,* Y_{sim} = *Mean).* We pursued the GWAS analysis and evaluated its performance with GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*. Second, for evaluating GWAS performance based on simulated matrix, we simulated QTL on top of different genotypic marker scores to create vectors of phenotypic values (Y_{sim} = *NImp*, Y_{sim} = *MVN-EM*, Y_{sim} = *Mean*). Then, we performed the GWAS analysis with those matrices as *GMat* (i.e. *NImp*, *MVN-EM* and *Mean*) and evaluated the GWAS performance. Finally, to evaluate imputation performance in a real phenotype dataset, we pursued GWAS analysis for wheat for four traits with high heritabilities. Details are explained in Methods section, and the general procedure is presented in Figure 1.

We summarized the results into power (*PO)* and false positive rate (*FPR)*. For further details, we present as additional material the results considering different thresholds.

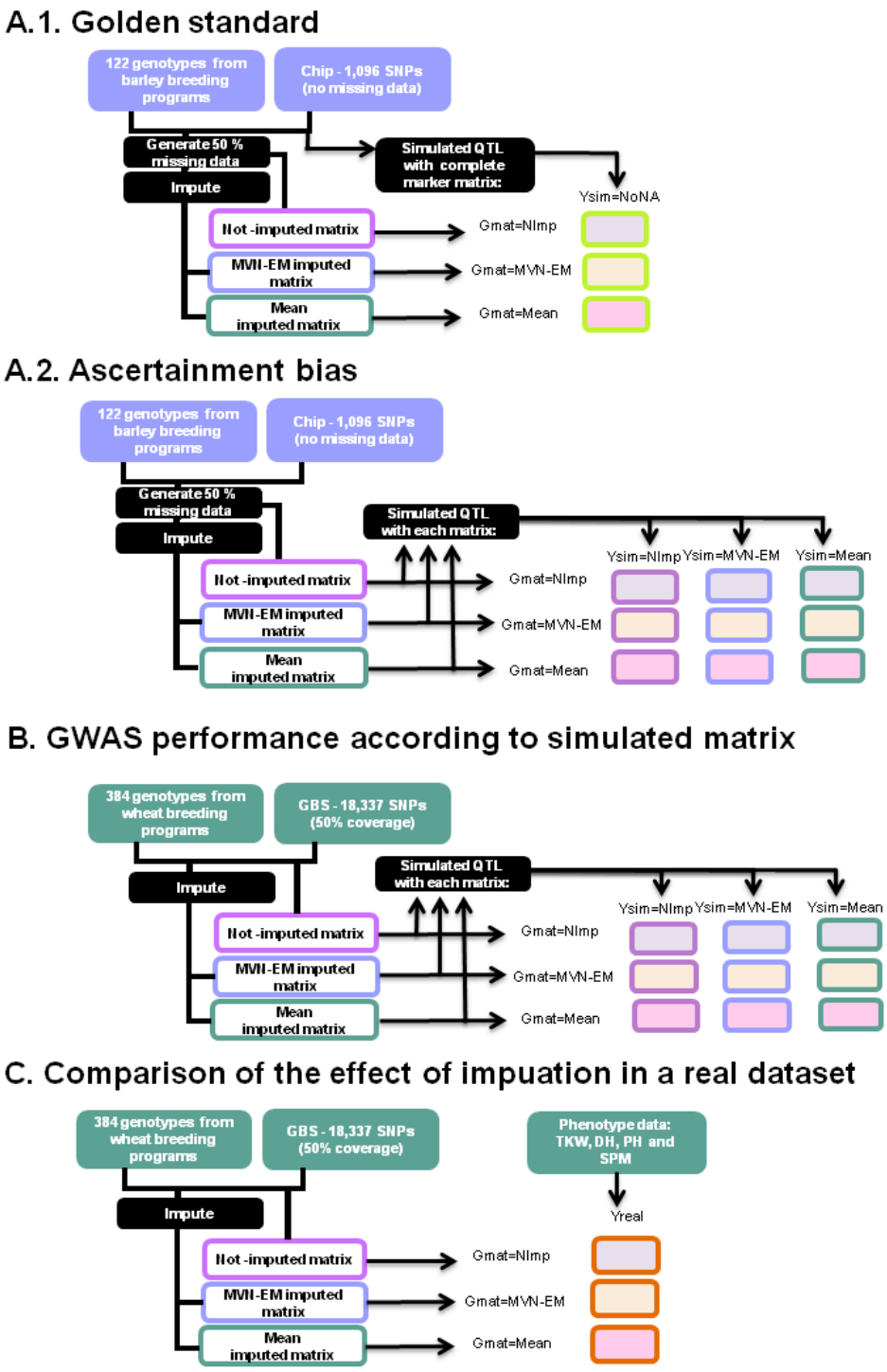


Figure 1 General scheme of the procedures we followed for each component. A. Procedures for golden standard (A.1) and ascertainment bias (A.2); B. Procedure for GWAS performance based on simulated matrix; C. Procedure for comparison of the

effect of imputation in a real phenotype dataset. Each procedure details the germplasm, genotypic and phenotypic dataset used, as well as simulation approach to obtain each phenotype vector and GWAS analysis marker scores matrices used. Procedures that used wheat data are in green and procedures that used barley data are in purple. DH, Days to Heading; GBS, Genotype-by-sequencing; MVN-EM, Multivariate Normal Expectation-Maximization; NImp, Not-imputed marker scores matrix; NoNA, No missing data marker scores matrix; PH, Plant Height; QTL, Quantitative Trait Loci; SNPs, Single-Nucleotide Polymorphism; SPM, Spikes Per Square Meter; TKW, Thousands Kernel Weight.

2.3.1. Ascertainment bias in imputation performance comparison

To evaluate the effect of imputation methods in GWAS analysis when no reference panel is available and there is a large proportion of missing data, we used a golden standard (i.e. a complete dataset). We found that simulating QTL over a complete dataset (Y_{sim} = *NoNA*), generating missing data, imputing with different methods, and then performing GWAS analysis (for general approach see Figure 1A.1), resulted in highest *PO* with *NImp* method for major QTL (GMat = *NImp*) and *MVN-EM* and *NImp* methods for minor QTL (GMat = *MVN-EM,* GMat = *NImp*), and smallest *FPR* with the *Mean* method (GMat = *Mean*, Figure 2), for different number of QTL (i.e. q=25 and q=50, data not shown) and heritabilities (i.e. h^2=0.2, h^2=0.4, h^2=0.6, h^2=0.7, h^2=0.9). Differences between *PO* were more evident for major QTL, but resulting in a small *PO* even for high heritability with a value of 0.3 (h^2=0.9, Figure 2). The highest values of *FPR* found with GMat = *MVN-EM* were also low (0.015, Figure 2). The same pattern was found when using different threshold levels (i.e. Bonferroni corrected by the effective number of independent markers –Li&Ji-, Figure 2, Bonferroni Additional file 1, and α = *0.01*, Additional file 2).

Furthermore, to evaluate if there are differences between imputation methods when a complete marker scores matrix is not available for the GWAS analysis, we simulated data with the different matrices, we created missing data on the complete marker scores matrix before simulating the QTL and compared the imputation performance (for general approach see Figure 1A.2). Using a phenotypic vector from

QTL simulated on top of the imputed marker scores matrices (i.e. Y_{sim} = *MVN-EM* or Y_{sim} = *Mean*), resulted in highest *PO* with both (GMat = *MVN-EM*, GMat = *Mean*) and smallest *FPR* with the *Mean* and *NImp* methods (GMat = *Mean,* GMat = *NImp*, Figure 3). However, when using a phenotypic vector from QTL simulated on top of raw, not-imputed marker scores (Y_{sim} = *NImp*) and evaluating imputation performance, resulted in highest *PO* with *MVN-EM* and *NImp* methods for major QTL and with the *MVN-EM* method for minor QTL, and smallest *FPR* with the *Mean* and *NImp* methods (GMat = *Mean,* GMat = *NImp*, Figure 3). This pattern was found for the different number of QTL (i.e. q=25 and q=50, data not shown) and different heritabilities (i.e. h^2=0.2, h^2=0.4, h^2=0.6, h^2=0.7, h^2=0.9, Figure 3). Differences between *PO* were also more evident for major QTL (Figure 3) as in the barley golden standard. The highest values of *FPR* found with GMat = *MVN-EM* was more evidenced when simulating with *NImp* matrix, but it was also low (0.015, Figure 3). Additionally, the same pattern was found using the different threshold levels (i.e. Bonferroni corrected by the effective number of independent markers – Li&Ji-, Figure 3, Bonferroni*,* Additional file 3, and = *0.01*, Additional file 4).

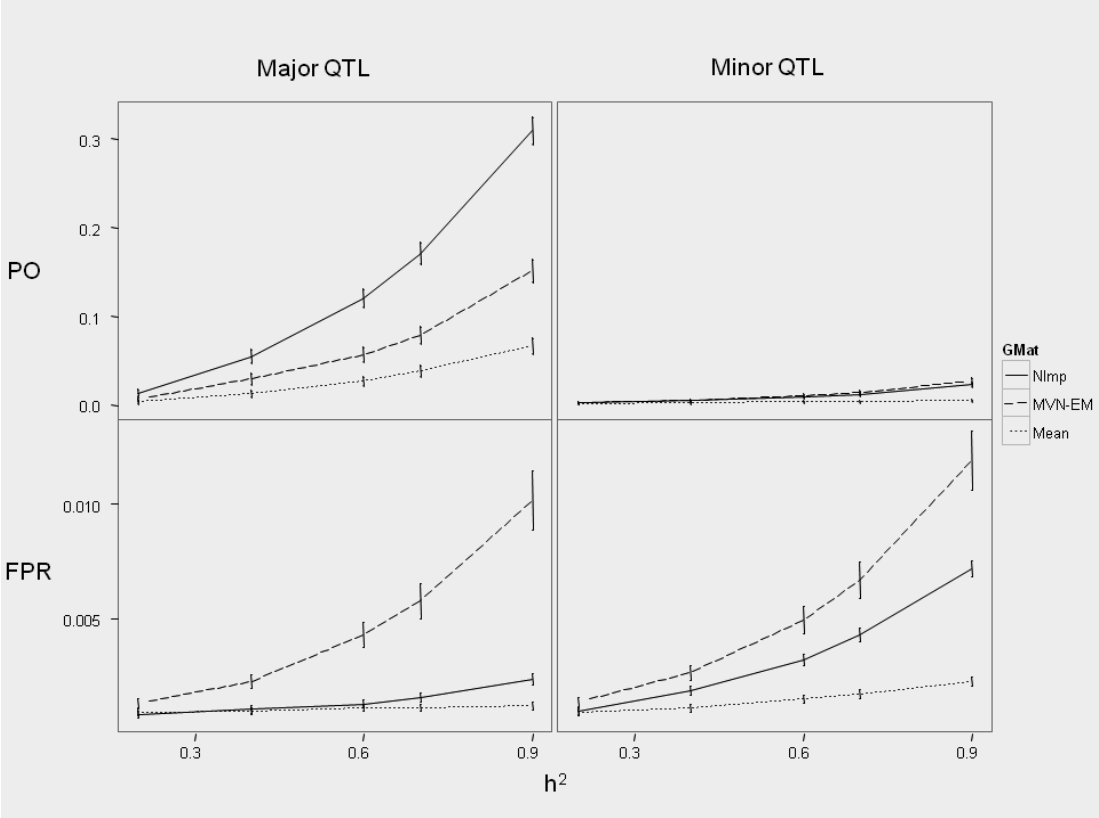


Figure 2 Power (*PO*) and false positives rate (*FPR*) for major and minor QTL with 25 QTL, for the golden standard from barley with a Bonferroni threshold corrected by the effective number of independent markers. Each parameter was calculated for the combinations of: heritabilties (h^2), a marker scores matrix to simulate the QTL (i.e. Y_{sim} = *NoNA*), and marker scores matrices to perform the GWAS analysis (i.e. GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*).

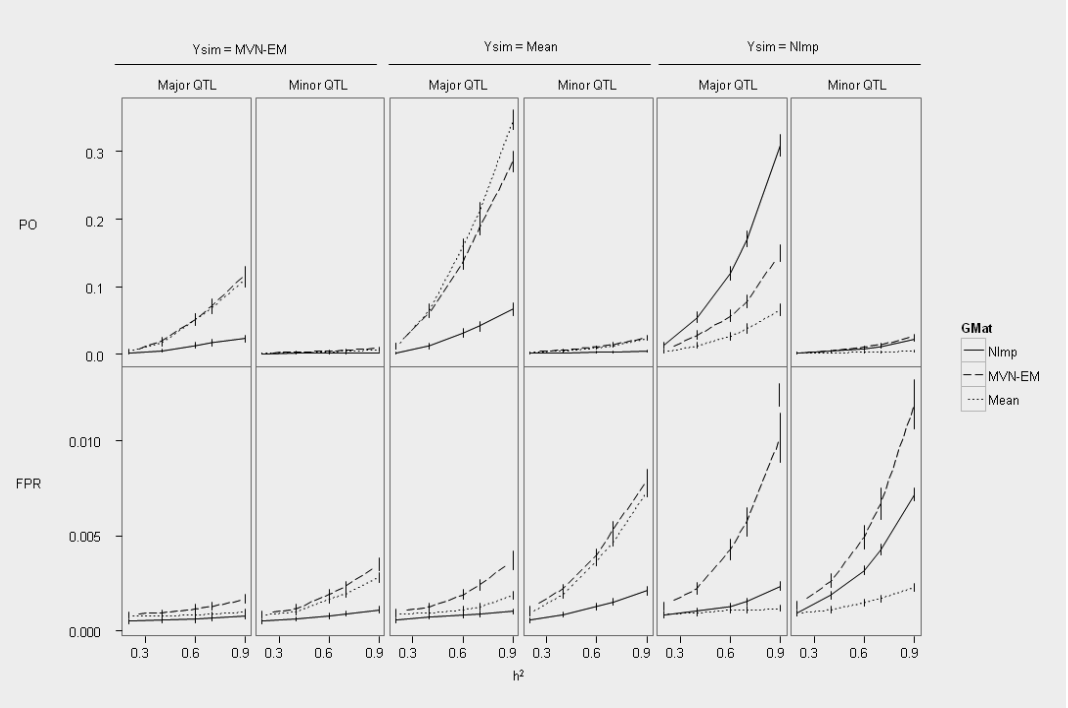


Figure 3 Power (*PO*) and false positives rate (*FPR*) with 25 QTL, for major and minor QTL for ascertainment bias in imputation performance comparison in barley, with a Bonferroni threshold corrected by the effective number of independent markers. Each parameter was calculated for the combinations of: heritabilties (h^2), marker scores matrices to simulate the QTL (i.e. Y_{sim} = *NImp*, Y_{sim} = *MVN-EM* and Y_{sim} = *Mean*), and marker scores matrices to perform the GWAS analysis (i.e. GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*).

2.3.2. GWAS performance based on simulated matrix

To evaluate the effect of the imputation methods on a GWAS performance using GBS wheat panel with 50% missing data (for general approach see Figure 1B), we imputed the GBS panel with both methods, simulated with each maker scores matrix (Y_{sim} = *NImp*, Y_{sim} = *MVN-EM* and Y_{sim} = *Mean*) and performed GWAS with each matrix (GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*). We detected the same pattern as the previous section (Figure 4), for the different number of QTL (i.e. q=25 and q=50, data not shown) and heritabilities (i.e. h^2=0.2, h^2=0.4, h^2=0.6, h^2=0.7, h^2=0.9, Figure 4). Differences between *PO* were more evident for major QTL, resulting in a *PO* for high heritability and major QTL of 0.75 (h^2=0.9, Figure

6). The highest values of *FPR* were found with simulating with the Y_{sim} = *NImp* and GMat = *MVN-EM* (Figure 4). Additionally, the same pattern was found using different threshold levels (i.e. Bonferroni corrected by the effective number of independent markers –Li&Ji-, Figure 4, Bonferroni, Additional file 5, and α = *0.01*, Additional file 6).

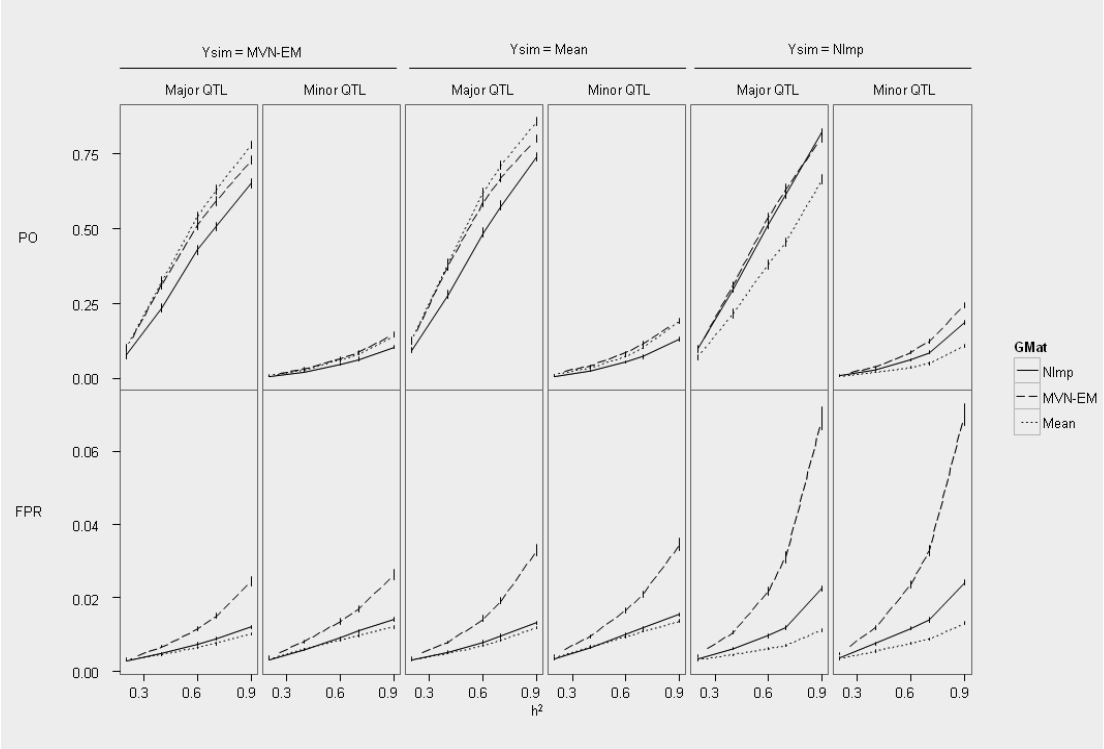


Figure 4 Power (*PO*) and false positives rate (*FPR*) with 25 QTL, for major and minor QTL to evaluate the GWAS performance based on simulated matrix with a Bonferroni threshold corrected by the effective number of independent markers. Each parameter was calculated for the combinations of: heritabilties (h^2), marker scores matrices to simulate the QTL (i.e. Y_{sim} = *NImp*, Y_{sim} = *MVN-EM* and Y_{sim} = *Mean*), and marker scores matrices to perform the GWAS analysis (i.e. GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*).

2.3.3. Comparison of the effect of imputation in a real dataset

We compared the QTL obtained for GWAS analysis using real phenotype data from wheat, with not imputed matrix (GMat = *NImp*), imputed with MVN-EM

method [4] matrix (GMat = *MVN-EM*) and imputed by the mean matrix (GMat = *Mean*). We considered the Bonferroni corrected by the effective number of independent markers threshold for multiple testing correction (Figure 5, Figure 6). For the four traits, plant height (PH, cm), days to heading (DH, days), thousand kernel weight (TKW, g) and spikes per square meter (SPM, number), we found different QTL when using imputed or not-imputed matrices (Figure 5). In general, *NImp* and *MVN-EM* matrices performed similar, having some QTL being detected by both methods (Figure 5). However, each matrix found unique QTL (Figure 5). For TKW, of the five QTL detected by the *NImp* matrix, four were detected with the *MVN-EM* matrix and two with the *Mean* matrix. For DH, of the five QTL detected by the *NImp* matrix, two were detected with the *MVN-EM* matrix and one was detected with the *Mean* matrix. Considering PH, of the two QTL found by the *NImp* matrix, one was detected with the *MVN-EM* matrix and no coincident QTL were found between the *NImp* matrix and the *Mean* matrix. Finally, for SPM, of the four QTL detected by the *NImp* matrix, one was detected with the *MVN-EM* matrix and two were detected with the *Mean* matrix.

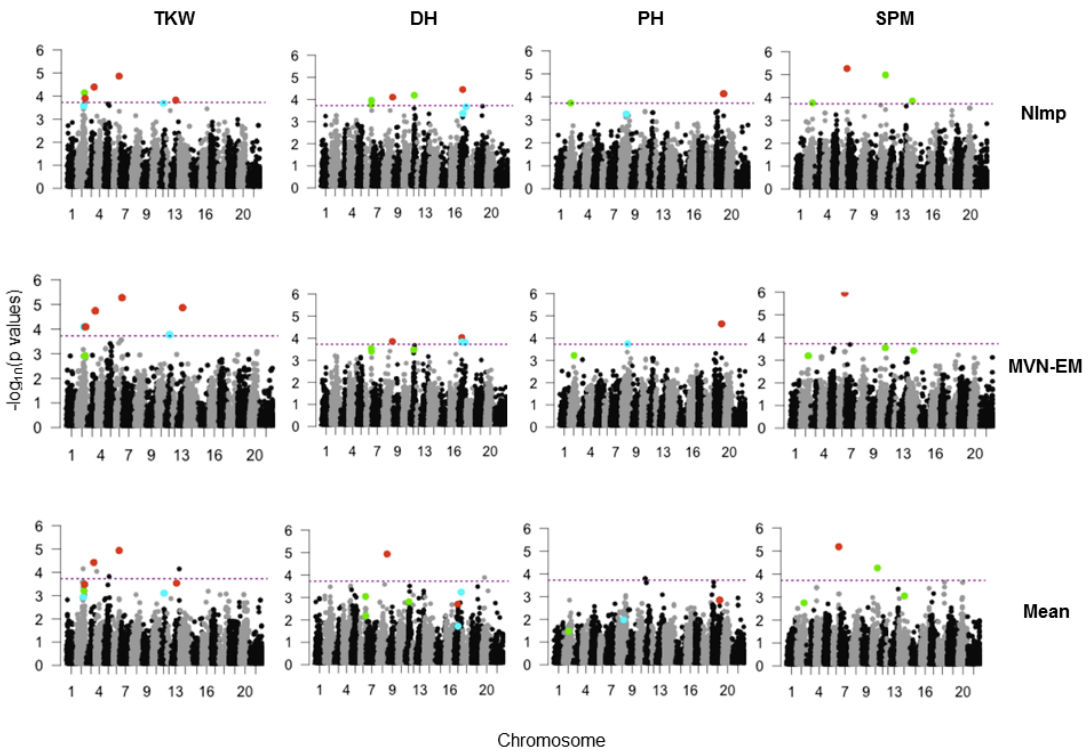


Figure 5 Manhattan plots of the GWAS analysis for real phenotype wheat data. For each trait evaluated a manhattan plot of the GWAS analysis is presented for each of *NImp* (not imputed), *Mean* (mean imputed) and *MVN-EM* (Multivariate Normal Expectation Maximization method) matrix. The phenotype traits are: DH, days to heading; PH, Plant Height; SPM, Spikes Per Square Meter; TKW, Thousands Kernel Weight. QTL detected exclusively by the *NImp* matrix are in green, QTL detected exclusively by the *MVN-EM* matrix are in skyblue and QTL detected by both matrices are in red.

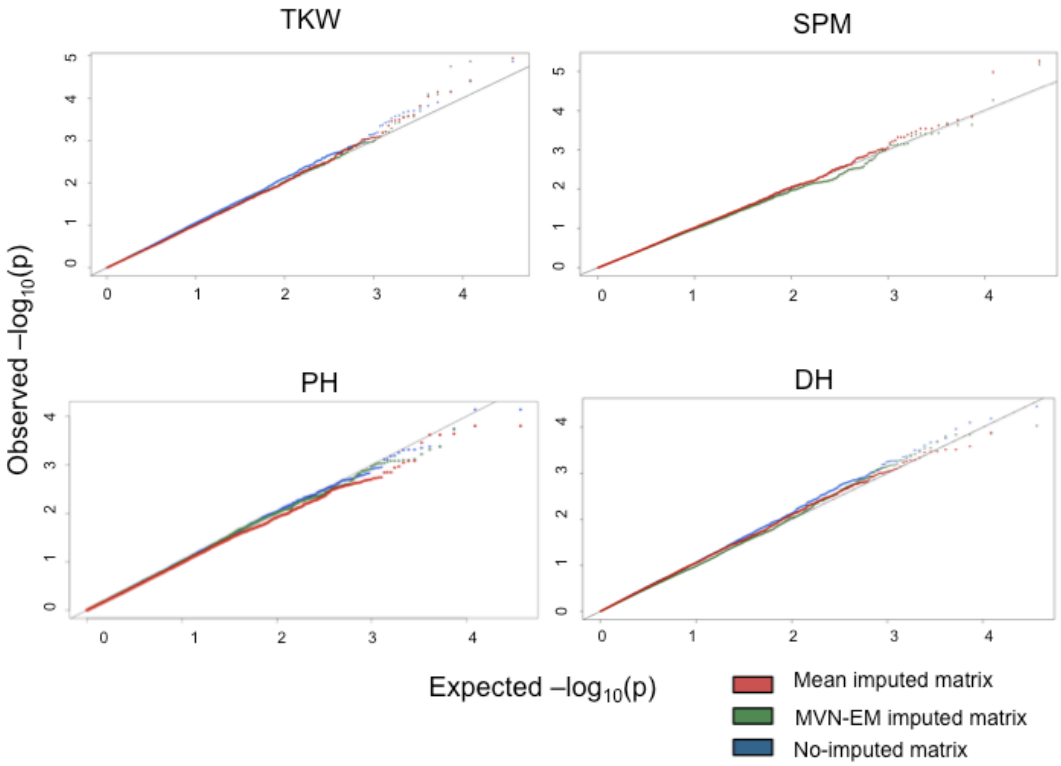


Figure 6 QQ plots of the p-values resulted form the GWAS analysis from real phenotype wheat data. For each trait measured and each marker scores matrix evaluated, a qq-plot of the p-values resulted form the GWAS analysis is presented. The marker scores matrices were: *NImp* (not imputed), *Mean* (mean imputed) and *MVN-EM* (Multivariate Normal Expectation Maximization method). The phenotype traits are: DH, days to heading; PH, Plant Height; SPM, Spikes Per Square Meter; TKW, Thousands Kernel Weight.

2.4. DISCUSSION

New whole-genome genotyping techniques are being developed and used for the diverse genetic analyses, like GWAS studies [3]. Although GBS is a powerful tool for genotyping hundreds of individuals with thousands of SNPs, it generates large amounts of missing data. Researchers have developed several strategies to impute missing data [8–11]. However, in GWAS analysis, imputation should be carefully evaluated because studies found that quality control of the data is an important source of loss of power [21], and we found an ascertainment bias in imputation evaluation.

2.4.1. Ascertainment bias in imputation performance comparison

When we used the barley golden standard marker scores matrix, the not-imputed marker scores matrix outperformed the imputation methods for all the combinations of the different parameters (Figure 2, Additional file 1, Additional file 2). The highest values of *FPR* found with the *MVN-EM* matrix and lowest values of *PO* found with *Mean* matrix, for all thresholds, could be a consequence of an imputation error affecting the signal of the QTL.

The fact that we found the same pattern when we artificially generated the missing data and when we used GBS data for all the combinations of parameters (Figure 4, Figure 6), gives the idea that there is an ascertainment bias. This ascertainment bias could be generated when there is no reference panel; the uncertainty of genotypic probability distributions due to the imputation is not incorporate in the GWAS analysis. Consequently, methods based on LD had found that if some restrictions are taking into account (i.e. strong LD among markers, low minor MAF, high genome coverage, and small population structure), the imputation accuracy is increased and hence the GWAS performance improved [16, 22].

The low *PO* detected for the barley marker scores matrix could be due to low LD between markers in the same LD blocks, because when there are unlinked QTL controlling a trait, the power is moderate even with large populations and high heritabilities [23]. Nevertheless, we do not expect unlinked QTL within the LD blocks because the LD blocks were defined by a single linkage agglomerative procedure [24], and because the genome coverage of the markers was very high, having 50% of its SNPs, at a distance smaller than 0.625 cm (Table 1). Therefore, we believe that the small population (122 lines) we used for this dataset could be affecting the *PO*, as the *PO* is a function of the population size [25].

The great differences found in *PO* and *FPR* between major and minor QTL, could be indicating that major QTL are the QTL mostly detected by any of the imputation methods.

Table 1. SNPs coverage on the golden standard matrix (i.e. complete SNP array), indicating for each chromosome (Chr = chromosome), the number of SNPs, the length (in cM), the largest gap without markers (cM), the median distance between pairs of adjacent markers, and the 25% and 75% quantiles of the adjacent marker distances.

Chr	SNPs number	Length (cM)	Largest gap (cM)	Median (cM)
1	125	139.78	10.74	0.625
2	187	150.27	8.21	0.58
3	178	170.88	6.59	0.58
4	131	121.65	7.5	0.6
5	201	194.03	8.05	0.57
6	147	129.38	8.62	0.47
7	127	166.56	10.53	0.49

2.4.2. GWAS performance based on simulated matrix

Differences were found when we simulated QTL on top of imputed marker scores and we evaluated the imputation performance (Figure 4). The performance of the GWAS analysis with the different methods (imputed or not-imputed matrices) changed. This is probably due to the imputation method use and the simulation. Other studies found that imputing with a reference panel improved precision [26], as we not had a reference panel, not-imputing was the best option for evaluating one marker at a time in GWAS analysis, specially for detecting major QTL as in the previous section.

2.4.3. Comparison of the effect of imputation in a real dataset

The traits evaluated in this paper were selected for having high heritability values and being related or a component of grain yield. The high heritability values may have reduced the differences between the QTL found with the *NImp* and *MVN-EM* matrices.

We found QTL where previous QTL were reported. The QTL found for TKW (chromosome 1B, bin 224 and 242) with the *NImp* and *MVN-EM* marker scores, and SPM (chromosome 1B, bin 224) with the *NImp* marker scores, are partially coincident with a QTL reported for green leaf area [27], a QTL reported for Near Differential Vegetative Index [28] and a QTL reported for yield, anthesis and plant height [31]. A QTL found for TKW (chromosome 1D, bin 205) with the *NImp* marker scores is coincident with a QTL reported for grain yield and plant height [29]. The QTL found for TKW (chromosome 2D, bin 167) with the three *GMat* marker scores, DH (chromosome 2D, bin 172) with the *NImp* marker scores, and SPM (chromosome 2D, bin 167) with the *NImp* marker scores, are coincident with a QTL reported for kernel weight, Near Differential Vegetative Index and flag leaf [27]. A QTL found for DH (chromosome 3B, bin 282) with the three *GMat* marker scores, is coincident with a QTL reported for grain filling duration [27]. A QTL found for SPM (chromosome 4A, bin 179) with the *NImp* and *MVN-EM* marker scores is coincident with a QTL reported for anthesis and plant height [29]. The QTL found for DH (chromosome 4B, bin 106) with the *NImp* marker scores is coincident with a QTL reported for yield and plant height [29]. A QTL found for TKW (chromosome 5A, bin 148) with the *NImp* and *MVN-EM* marker scores is coincident with a QTL reported for yield, anthesis and plant height [29]. A QTL found for SPM (chromosome 5B, bin 173) with the *NImp* marker scores is coincident with a QTL reported for yield and plant height [29]. A QTL found for DH (chromosome 6B, bin 116) with the *NImp* and *MVN-EM* marker scores, is coincident with a QTL for yield and plant height [29]. A QTL found for PH (chromosome 7A, bin 225) with the *NImp* and *MVN-EM* marker scores, is coincident with yield and anthesis [29]. These positions are based on bins and should be regarded as an approximation. These could be improved after the draft of the genome is available [30].

As we found that QTL detected by the *NImp* marker scores matrix and the *MVN-EM* imputed marker scores matrix were similar, we believe that imputation should not be taking into account because no improvement is being detected.

2.5. CONCLUSION

Imputation can introduce an ascertainment bias to GWAS analysis. Comparing the GWAS performance by the power (*PO*) and false positive rate (*FPR*) with imputed or not-imputed marker scores matrices when we performed the simulations, poorer performance was found when an imputed marker scores matrix was used. Additionally, the *PO* and *FPR* changed in a clear way between major and minor QTL, showing that differences among imputation methods were more evident for major QTL and that the detection of minor QTL is negligible. Thus, although imputation can improve the performance of certain analysis like GS, when GWAS analysis is performed imputation by the mean or with the MVN-EM method is not encouraged.

2.6. METHODS

2.6.1. Dataset

We used three datasets: (1) a complete SNPs barley panel array, and (2) a GBS wheat marker scores matrix with an average of 50% coverage and phenotypic data (for general approach see Figure 1).

The complete barley SNP marker scores array dataset consisted in a panel of 122 barley advanced inbred lines from a population of 360 described in [31]. Briefly, 1,096 SNPs from the Barley Oligonucleotide Pool Assay-1 (BOPA 1) were selected [32, 33]. For further details of dataset see [31]. The 122 lines were selected to form two complete datasets, without missing information.

The wheat GBS dataset was a panel of 384 advanced inbred lines from breeding programs: 186 genotypes from the National Wheat Breeding Program from Uruguay (INIA-Uruguay, Instituto Nacional de Investigación Agropecuaria), 55 genotypes from the National Wheat Breeding Program from Chile (INIA-Chile), and 143 genotypes from the International Breeding Center of Maize and Wheat

(CIMMYT, Centro Internacional de Mejoramiento de Maíz y Trigo). The CIMMYT genotypes share common ancestors with the INIA-Chile genotypes (see [34] for more details).

DNA was extracted by the DNeasy Plant Maxi Kit (QIAGEN). Library construction was conducted in Kansas State University (Manhattan, Kansas) using a PstI-MspI GBS protocol [4]. The sequencing was performed on an Illumina Hi-Sequation 2000 at the DNA core facility at the University of Missouri, Columbia, Missouri, and the McGill Univesity-Génome Quebec Innovation Centre (Montreal, Canada) for each set of libraries. SNPs (Single-Nucleotide Polymorphism) were obtained using the Tassel-GBS Pipeline [35]. The base quality and distribution of sequences was studied with the Galaxy (http://galaxy.psu.edu/) software. SNPs with less than 50 % coverage and with minor allele frequency (MAF) smaller than 10% were excluded. Sequences were blasted to the SyntheticxOpata map (synop) using the blastn function from NCBI-BLAST+ package using the number of descriptions and the number of threads set to one. Therefore, SNPS were placed into recombination bins, defined by each observed recombination across the population, where for all markers within a bin, the alleles received by a line should have originated from the same parent [5]. A final matrix set of 18,337 SNPs was obtained (Table 2), with a median distance between markers for all chromosomes of cero due to the use of a bin map.

The phenotypic data was obtained from an evaluation in a Mediterranean environment in Santa Rosa-Chile in 2011 (36º 329' S, 71º 559' W; 217 m.a.s.l.). The field was irrigated with 50 mm m^{-2} at each of four moments: tillering, flag leaf emergence, heading date, and grain filling (see [34] for further details). The experimental design was an alpha-lattice with 20 replications and 20 incomplete blocks. The traits evaluated were: plant height (PH, cm) evaluated from the base of the plant to the flower insertion, days to heading (DH, days) was recorded when 50% of the culms showed emerged ears, thousands kernel weight (TKW, g), and spikes per square meter (SPM, number). We obtained the best linear unbiased predictors (BLUPs) using the following model for each trait: $y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{k(j)} + \varepsilon_{ijk}$ where y_{ijk} is the value for the phenotypic trait corresponding to the *i-th* genotype, *j-th*

replication, and *k-th* incomplete block, μ is the overall mean, α_i is the random effect of the *i-th* genotype with $\alpha_i \sim N(0, \sigma^2_\alpha)$, β_j is the effect of the *j-th* replication, $\delta_{k(j)}$ is the random effect of the *k-th* incomplete block within the *j-th* replication with $\delta_{k(j)} \sim N(0, \sigma^2_\delta)$, ε_{ijk} is the experimental error corresponding to the *i-th* genotype, *j-th* replication and *k-th* incomplete block with $\varepsilon_{ijk} \sim N(0, \sigma^2_\varepsilon)$.Genotypic means were estimated with the function *lmer* (*lme4* package) in R statistical software [36]. Broad sense heritabilities were estimated in R statistical software [36] using the above model (Table 3).

Table 2. SNPs coverage on the GBS genotypic matrix, indicating for each chromosome (Chr = chromosome), the number of SNPs, the length (in cM) and the largest gap without markers (cM).

Chr	SNPs number	Length (cM)	Largest gap (cM)	Median (cM)
1A	821	266	33	0
1	1282	294	22	0
1D	255	242	25	0
2A	900	242	22	0
2B	1746	266	38	0
2D	327	182	27	0
3	929	329	28	0
3B	1912	290	30	0
3D	270	287	29	0
4A	907	234	28	0
4B	610	177	31	0
4D	74	130	45	0
5A	1023	232	26	0
5B	1270	316	22	0
5D	197	306	29	0
6A	883	237	34	0
6B	1302	232	25	0
6D	243	276	28	0
7A	1456	323	24	0
7B	1660	263	40	0
7D	270	337	45	0

Table 3. Broad sense heritability (h^2) for the real wheat panel for all traits in Santa Rosa- Chile 2011.

Trait	Santa Rosa- Chile 2011
Plant height (cm)	0.78
Days to heading (days)	0.97
Thousand kernel weight (g)	0.93
Spikes per square meter (number)	0.76

2.6.2. Imputation methods

For the barley SNP array panel, we started with a genotype by marker scores matrix with 122 genotypes (rows) and 1,096 markers (columns) without missing values Markers were scored as the number of alleles {1, -1}. Then, we randomly generated missing values in order to have the same coverage as the GBS panel (50%). Finally, two methods were used to fill in those missing values, MVN-EM which considers the realized additive relationship matrix between the lines and an EM approach assuming that marker genotypes follow a multivariate normal distribution [4] and the Mean score per marker (i.e. the expected allele count at the particular marker). Imputation was conducted in R statistical software [36] with *A.mat* function (*rrBLUP* package).

For the wheat GBS panel, we started with a genotype by marker scores matrix with 384 genotypes (rows) and 18,337 markers (columns) with 50% of missing values. Markers were scored as the number of alleles {NA, 1, -1}. We used the same methods as the previous sections to impute by MVN-EM and Mean..

2.6.3. Simulation procedure

To evaluate the effect of imputation using a golden standard with the barley SNP array, we created phenotypic vectors simulating QTL on top of the complete barley marker scores matrix ($Y_{sim} = NoNA$). The phenotypic vectors were the sum of the effects of genotypic and residual terms, $Y_{sim} = g + e$. The genotypic effect was calculated as the sum of the markers (selected as QTL) effects and markers effects were obtained from a Beta(2, 6) distribution. Markers selected as QTL were obtained

form the LD blocks defined from a single linkage agglomerative procedure [24] with euclidean distances between markers and a minimum of 1.5 cM to consider independent groups. QTL with major effect were defined as the above the 75%, and QTL with minor effect was defined as the below or equal the 75%. The residual term was obtained by sampling from N(0, ^2_e), where $^2_e = (1- h^2)\ ^2_g/ h^2$ and ^2_g was the variance of the realized *g*. One vector for the combinations of number of QTL (i.e. *q*=25 and *q*=50), different heritabilities (i.e. h^2=0.2, h^2=0.4, h^2=0.6, h^2=0.7, h^2=0.9), and for each one of 500 iterations was created. Then, we created missing data at random, imputed (i.e. GMat =*NImp*, GMat = *MVN-EM* and GMat = *Mean*) and pursued the GWAS analysis with each combination of genotypic matrix, evaluating *PO* and *FPR* (for general approach see Figure 1A.1). Additionally, for the ascertainment bias evaluation, we first created the missing data and then simulated the QTL on top of each matrix: Y_{sim} = *NImp* for the not-imputed marker scores, Y_{sim} = *MVN-EM* for the imputed with MVN-EM [4] marker scores, Y_{sim} = *Mean* for the imputed by the mean marker scores. Finally, we performed the GWAS analysis with each genotypic marker scores (i.e. GMat =*NImp*, GMat = *MVN-EM* and GMat = *Mean*) and for each phenotypic vector (i.e. Y_{sim} = *NImp,* Y_{sim} = *MVN-EM* and Y_{sim} = *Mean*, for general approach see figure 1A.2). We therefore compared the *PO* and *FPR*.

For evaluating GWAS performance based on simulated matrix with the wheat GBS panel data, we created vectors of phenotypic values (i.e. Y_{sim} = *NImp,* , Y_{sim} = *MVN-EM* and Y_{sim} = *Mean*). Each phenotypic vector was simulated for different number of QTL (i.e. *q*=25 and *q*=50) and different heritabilities (i.e. h^2=0.2, h^2=0.4, h^2=0.6, h^2=0.7, h^2=0.9) as in the previous section. In order to avoid collinearity, LD blocks were defined as the bins in each chromosome and a marker chosen at random within each LD block was considered a QTL. One vector for each combination of the parameters and for each one of 500 iterations was created. We performed the simulations in R statistical software [36].

2.6.4. GWAS analysis

For the GWAS analysis, the mixed model described by [37] was used: $y = X\beta + Qv + \underline{Z}u + e$, where y is the phenotypic vector (n x 1) with *n* the total number of lines, X is a (n x m) SNPs matrix with *m* the number of SNPs coded as described before {NA, 1, -1}, β is a (m x 1) vector of allelic effects to be estimated, Q is a (n x q) incidence matrix with *q* origin's groups, v is a (n x 1) populations fixed effect vector, Z is the genotypic incidence matrix, u is the vector of random background polygenic effects, $u \sim N(0, A \sigma^2_g)$, where *A* is the realized additive relationship matrix obtained with the *A.mat* function from package *rrBLUP* in R statistical software [36] and e is the residual error, $e \sim N(0, \sigma^2_e)$. For each Y_{sim}, we used the three genotypic marker scores to recover the QTL, *GMat* (i.e. *NImp*, *MVN-EM* and *Mean*). We performed the analysis for three different thresholds (*threshold*) to define markers as significant: (1) Bonferroni, (2) Bonferroni correction using the effective number of markers, Li&Ji method [38], and (3) liberal threshold of $\alpha = 0.01$. GWAS analysis was accomplished with *GWAS* function from *rrBLUP* package in R statistical software [36]. We defined as true positives (TP) the number of bins with a QTL and at least one significant marker; false positives (FP) the number of bins with no QTL and at least one significant marker; true negatives (TN) the number of bins with no QTL and no significant markers, and false negatives (FN) the number of bins with QTL and no significant markers. We evaluated power (*PO*=TP/ (TP+FN)) and false positive rate (*FPR*=FP/ (FP+TN)) [39] for QTL detection. We evaluated performance for major and minor QTL detection.

2.7. REFERENCES

1. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12:499–510.
2. Heslot N, Rutkoski J, Poland J, Jannink J-L, Sorrells ME. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One. 2013;8:e74612.

3. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, McMullen MD, Holland JB, Buckler ES, Gardner CA. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 2013;14:R55.
4. Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink J-L. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Genome J. 2012;5:103.
5. Poland JA, Brown PJ, Sorrells ME, Jannink J-L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One. 2012;7:e32253.
6. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6:e19379.
7. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34:816–34.
8. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39:906–13.
9. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78:629–44.
10. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
11. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81:1084–97.

12. Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet. 2008;124:439–50.
13. Jannink J-L, Iwata H, Bhat PR, Chao S, Wenzl P, Muehlbauer GJ. Marker Imputation in Barley Association Studies. Plant Genome J. 2009;2:11.
14. Hao K, Chudin E, McElwee J, Schadt EE. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genet. 2009;10:27.
15. Pei Y-F, Li J, Zhang L, Papasian CJ, Deng H-W. Analyses and comparison of accuracy of different genotype imputation methods. PLoS One. 2008;3:e3551.
16. Iwata H, Jannink J-L. Marker Genotype Imputation in a Low-Marker-Density Panel with a High-Marker-Density Reference Panel. Accuracy Evaluation in Barley Breeding Lines. Crop Sci. 2010;50:1269.
17. Guan Y, Stephens M. Practical issues in imputation-based association mapping. PLoS Genet. 2008;4:e1000279.
18. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, Gupta N, Neale BM, Daly MJ, Sklar P, Sullivan PF, Bergen S, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Purcell SM, Haas DW, Liang L, Sunyaev S, Patterson N, de Bakker PIW, Reich D, Price AL. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012;44:631–5.
19. Almeida MAA, Oliveira PSL, Pereira T V, Krieger JE, Pereira AC: An empirical evaluation of imputation accuracy for association statistics reveals increased type-I error rates in genome-wide associations. BMC Genet. 2011;12:10.
20. Aulchenko YS, Struchalin M V, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. BMC Bioinformatics. 2010;11:134.
21. De Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet. 2008;17:122–28.
22. Jannink J-L, Iwata H, Bhat PR, Chao S, Wenzl P, Muehlbauer GJ. Marker Imputation in Barley Association Studies. Plant Genome J. 2009;2:11.

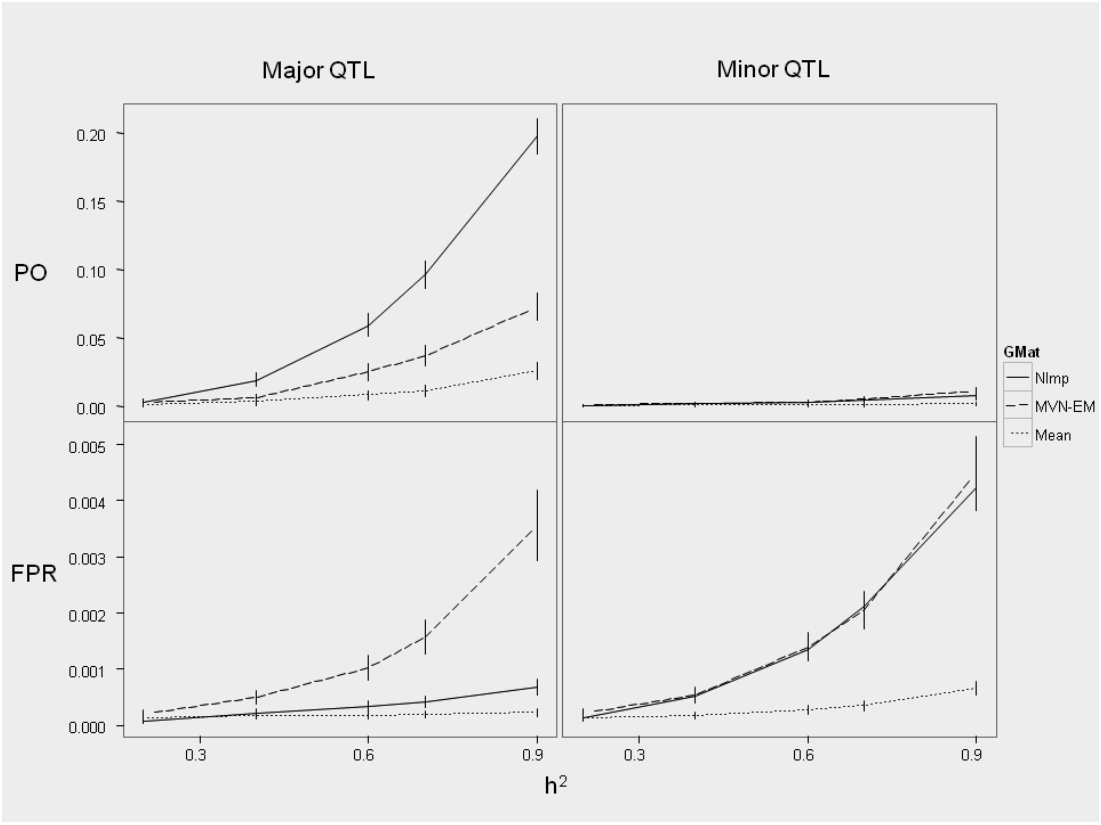
23. Bernardo R. Breeding for Quantitative Traits in Plants. 2nd ed. Minnesota: Stemma Press; 2010.
24. Sibson R. SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal 1973;30–34.
25. Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 1990;124:743–756.
26. He S, Zhao Y, Mette MF, Bothe R, Ebmeyer E, Sharbel TF, Reif JC, Jiang Y. Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L.). BMC Genomics. 2015;16:1–12.
27. Edae EA, Byrne PF, Haley SD, Lopes MS, Reynolds MP. Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. Theor Appl Genet. 2014;127:791–807.
28. Bennett D, Reynolds M, Mullan D, Izanloo A, Kuchel H, Langridge P, Schnurbusch T. Detection of two major grain yield QTL in bread wheat (Triticum aestivum L.) under heat, drought and high yield potential environments. Theor Appl Genet. 2012;125:1473–85.
29. Mathews KL, Malosetti M, Chapman S, McIntyre L, Reynolds M, Shorter R, van Eeuwijk F. Multi-environment QTL mixed models for drought stress adaptation in wheat. Theor Appl Genet. 2008;117:1077–91.
30. Mayer KFX, Rogers J, Dole el J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski AJ, Sourdille P, Endo TR, Kubalakova M, Ihalikova J, Dubska Z, Vrana J, Perkova R, Imkova H, Febrer M, Clissold L, McLay K, Singh K, Chhuneja P, Singh NK, Khurana J, Akhunov E, Choulet F, Alberti A, Barbe V, Wincker P, Kanamori H, et al. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science. 2014;345(6194):1251788.
31. Gutiérrez L, Germán S, Pereyra S, Hayes PM, Pérez CA, Capettini F, Locatelli A, Berberian NM, Falconi EE, Estrada R, Fros D, Gonza V, Altamirano H, Huerta-Espino J, Neyra E, Orjeda G, Sandoval-Islas S, Singh R, Turkington K, Castro AJ. Multi-environment multi-QTL association mapping identifies

disease resistance QTL in barley germplasm from Latin America. Theor Appl Genet. 2014:501–16.

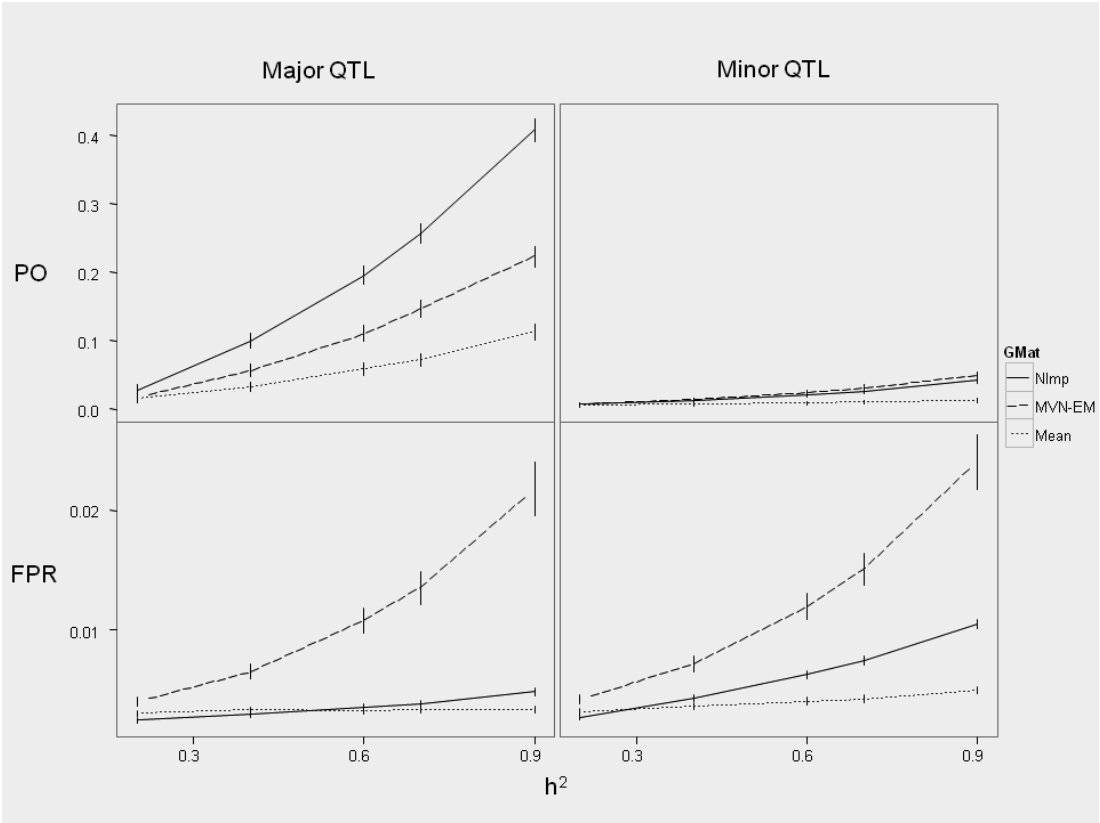
32. Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson JT, Wanamaker S, Bozdag S, Roose ML, Moscou MJ, Chao S, Varshney RK, Sz P, Sato K, Hayes PM, Matthews DE, Kleinhofs A, Muehlbauer GJ, Deyoung J, Marshall DF, Madishetty K, Fenton RD, Condamine P, Graner A, Waugh R. Development and implementation of high-throughput SNP genotyping in barley. 2009;13:1–13.
33. Szűcs P, Blake VC, Bhat PR, Chao S, Close TJ, Cuesta-Marcos A, Muehlbauer GJ, Ramsay L, Waugh R, Hayes PM. An Integrated Resource for Barley Linkage Map and Malting Quality QTL Alignment. Plant Genome J. 2009;2:134.
34. Lado B, Matus I, Rodríguez A, Inostroza L, Poland JA, Belzile F, del Pozo A, Quincke M, Castro M, von Zitzewitz J. Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3 (Bethesda). 2013;3:2105–14.
35. Glaubitz JC, Casstevens TN, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One. 2014;9(2):e90346.
36. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/ 2015.
37. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 2006;38:203–208.
38. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb). 2005;95:221–27.
39. Chengsong Z, Jianming Y. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics. 2009;182:875–88.

2.8. ADDITIONAL MATERIAL

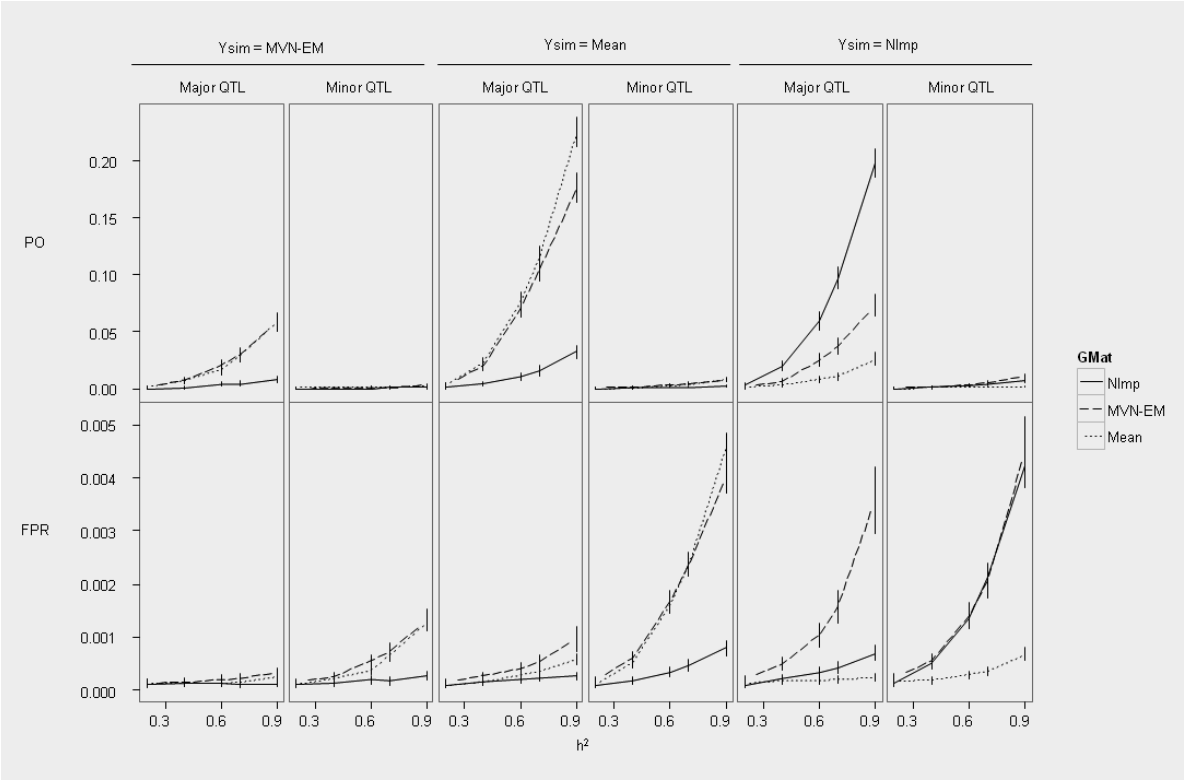
2.8.1. <u>Additional file 1</u>: Power (*PO*) and false positives rate (*FPR*) for major and minor QTL with 25 QTL, for the golden standard form barley, with a Bonferroni threshold. Each parameter was calculated for the combinations of: heritabilties (h^2), a marker scores matrix to simulate the QTL (i.e. Y_{sim} = *NoNA*), and marker scores matrices to perform the GWAS analysis (i.e. GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*).



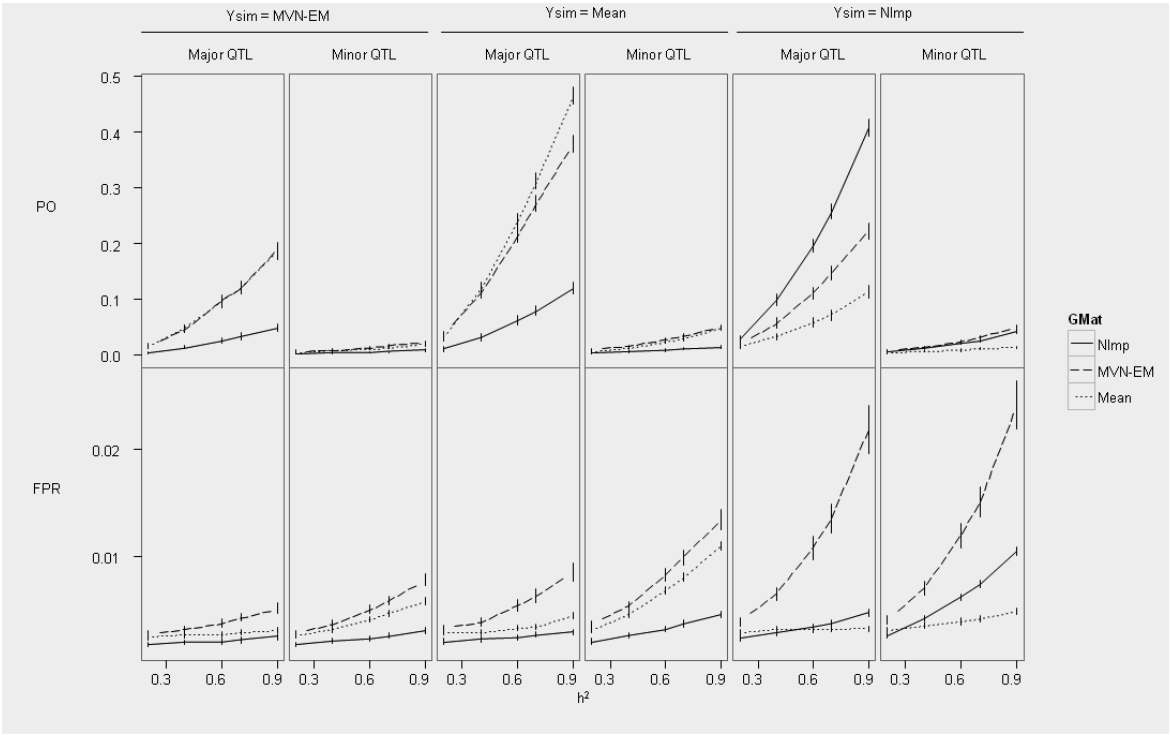
2.8.2. <u>Additional file 2</u>: Power (*PO*) and false positives rate (*FPR*) for major and minor QTL with 25 QTL, for the golden standard from barley, with $\alpha = 0.01$ **threshold.** Each parameter was calculated for the combinations of: number of QTL (q), heritabilties (h^2), a marker scores matrix to simulate the QTL (i.e. Y_{sim} = *NoNA*), and marker scores matrices to perform the GWAS analysis (i.e. GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*).



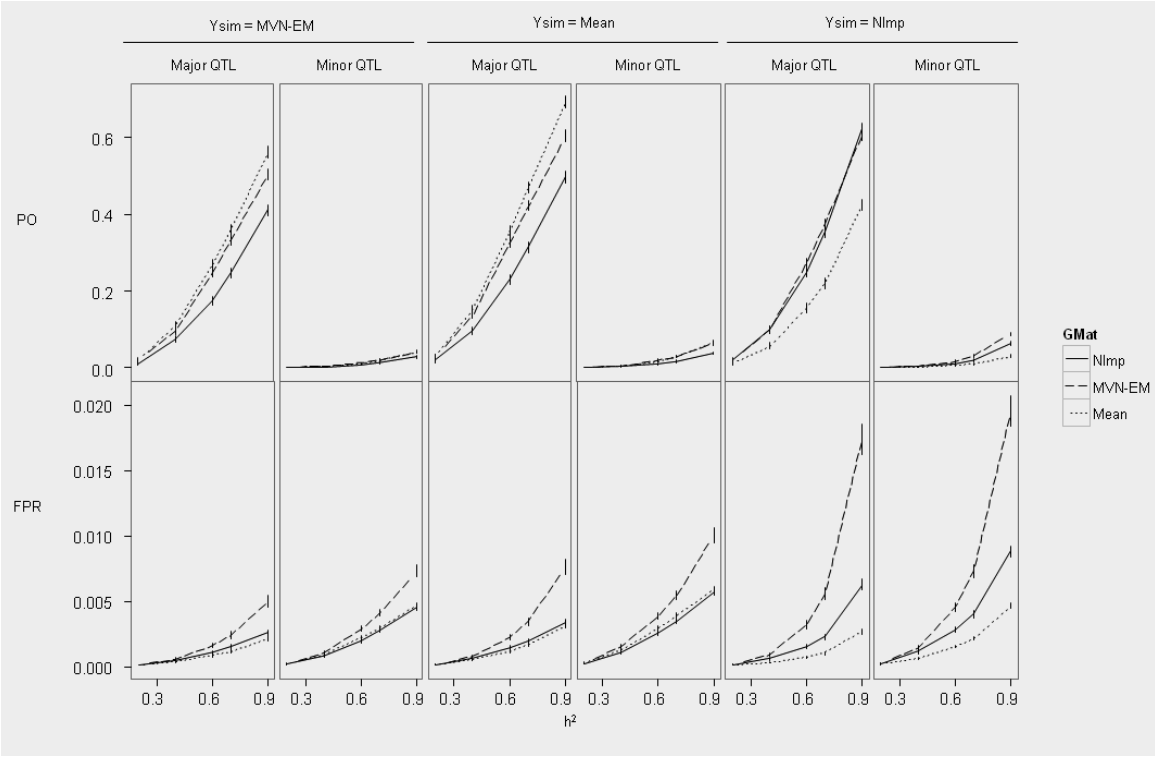
2.8.3. <u>Additional file 3</u>: Power (*PO*) and false positives rate (*FPR*) with 25 QTL, for major and minor QTL for ascertainment bias in imputation performance comparison in barley, with a Bonferroni threshold. Each parameter was calculated for the combinations of: heritabilties (h^2), marker scores matrices to simulate the QTL (i.e. Y_{sim} = *NImp*, Y_{sim} = *MVN-EM* and Y_{sim} = *Mean*), and marker scores matrices to perform the GWAS analysis (i.e. GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*).



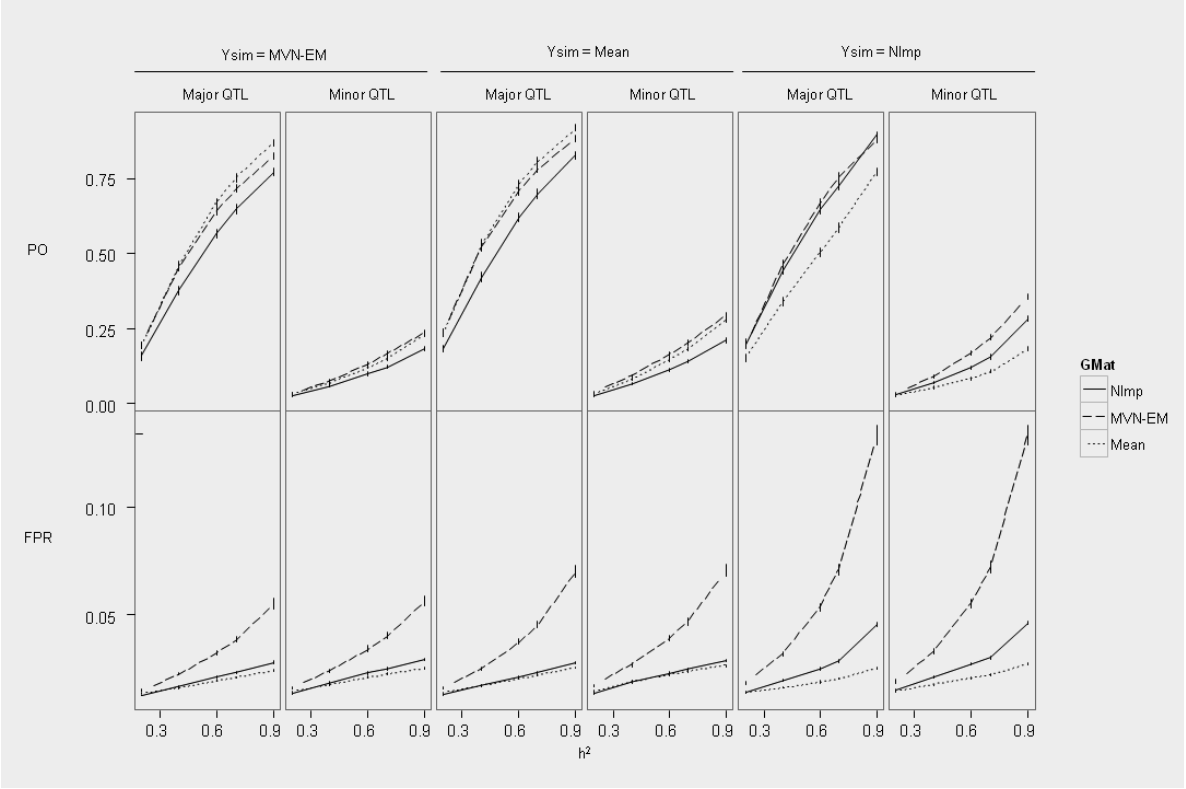
2.8.4. <u>Additional file 4</u>: Power (*PO*) and false positives rate (*FPR*) with 25 QTL, for major and minor QTL for ascertainment bias in imputation performance comparison in barley, with a = 0.01 threshold. Each parameter was calculated for the combinations of: heritabilties (h^2), marker scores matrices to simulate the QTL (i.e. Y_{sim} = *NImp*, Y_{sim} = *MVN-EM* and Y_{sim} = *Mean*), and marker scores matrices to perform the GWAS analysis (i.e. GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*).



2.8.5. <u>Additional file 5</u>: Power (*PO*) and false positives rate (*FPR*) with 25 QTL, for major and minor QTL to evaluate the GWAS performance based on simulated matrix with a Bonferroni threshold. Each parameter was calculated for the combinations of: heritabilties (h^2), marker scores matrices to simulate the QTL (i.e. Y_{sim} = *NImp*, Y_{sim} = *MVN-EM* and Y_{sim} = *Mean*), and marker scores matrices to perform the GWAS analysis (i.e. GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*).



2.8.6. <u>Additional file 6</u>: Power (*PO*) and false positives rate (*FPR*) with 25 QTL, for major and minor QTL to evaluate the GWAS performance based on simulated matrix with a = 0.01 threshold. Each parameter was calculated for the combinations of: heritabilties (h^2), marker scores matrices to simulate the QTL (i.e. Y_{sim} = *NImp*, Y_{sim} = *MVN-EM* and Y_{sim} = *Mean*), and marker scores matrices to perform the GWAS analysis (i.e. GMat = *NImp*, GMat = *MVN-EM* and GMat = *Mean*).



3. THE GENETIC BASIS FOR AGRONOMICAL AND PHYSIOLOGICAL TRAITS IN WHEAT UNCOVERED BY A MULTI-TRAIT MULTI-ENVIRONMENT STUDY[2]

3.1. ABSTRACT

Wheat (*Triticum aestivum* L.) is the third most important cereal crop of the world in terms of production. Understanding the genetic basis of wheat yield-related traits considering genotype by environment interaction can improve wheat productivity. The goal of this study was to tackle the understanding of grain yield by its components and related-traits, integrating information in a multi-trait and multi-environment Genome-Wide Association (GWAS) analysis. A set of 384 advanced genotypes of wheat from INIA-Uruguay, INIA-Chile and CIMMYT was evaluated in two Mediterranean environments in central Chile: Cauquenes under rainfed conditions in 2011 and Santa Rosa under mild water stress and fully irrigated, in 2011 and 2012. Sixteen physiological traits were evaluated and 28,217 markers were obtained. Heritabilities were medium to high for most traits. For the multi-trait GWAS approach, primarily main effect QTL (Quantitative Trait Loci) were found by studying correlated traits within subgroups (grain yield components, leaf related traits, and morphology and phenology traits), with the exception of two QTL by trait interactions (QTI) for morphology and phenology traits. Although within the subgroups not all traits were strong and positive correlated, main effect QTL found could indicate that the development of the crop is mostly affected by the same genetic basis for biology related traits. For the multi-environment GWAS approach, only main QTL were found, implying that QTL with positive effect for grain yield in one environment, could be introduced in the breeding program's populations and improve the genotypes performance in the other environments evaluated.

[2] Artículo a publicar en: The Plant Genome

3.2. INTRODUCTION

Wheat is the third most important crop in terms of total world production with 670 million ton produced in 2012 (FAOSTAT, 2014). However, food security could be compromised by the increase in food demand due to population growth (Mueller et al., 2012) and climate change (Ewert et al., 2005). Plant breeding has successfully increased wheat grain yield (Fischer, 2007) but the rate of yield gain has decreased in the last decades (Acreche et al., 2008; Reynolds et al., 2012; Bustos et al., 2013). Therefore, improving wheat productivity is key for responding to the increase in food demand and climate change while decreasing agriculture´s global environmental footprint (Mueller et al., 2012). Although genetic gains are still obtained for wheat, global wheat demand is growing at a faster rate than genetic gains (Barnabás et al., 2008; García et al., 2013). Therefore, new breeding strategies should be pursued (Fischer, 2007).

Yield (per se) is a challenging breeding target because it is a complex trait determined by many genes (Slafer and Araus, 2005; Alimi et al., 2012), with kernel weight, grains per spike and spikes per meter square as yield components (Kjaer and Jensen, 1996). Yield is also affected by different biotic factors like incidence and severity of diseases, especially *Puccinia graminis tritici, Puccinia tritici* (rust, Chen, 2005; Singh et al., 2008) and *Fusarium graminearum* (Windels, 2000). Agronomic traits like grain yield have strong genotype by environment interaction, GEI (Hayes et al., 1993; Boer et al., 2007; Mathews et al., 2008; van Eeuwijk et al., 2010; Malosetti et al., 2013; Alimi et al., 2013). Therefore, improving yield per se is a challenge (Alimi et al., 2012). Consequently, improving yield by studying the genetic basis of agronomic and physiological yield-related traits could provide a better understanding of the trait (Slafer and Araus, 2005; Fischer, 2007) leading to faster genetic gains.

GWAS is useful to analyze the genetic basis of quantitative traits by searching whole genome marker-trait associations (Zhu et al., 2008). Specifically, a multi-trait QTL analysis allows the simultaneous study of genetically correlated traits (Boer et al. 2007; Malosetti et al. 2007b; van Eeuwijk et al. 2010; Alimi et al. 2013; Malosetti et al. 2013; El-Soda et al. 2014). These approaches allow the discovery of pleiotropic

and closely linked QTLs, and improve the power of QTL mapping (Alimi et al., 2013). It has also helped in improving the selection of some primary traits with low heritabilities (Jiang and Zeng, 1995). Therefore, a multi-trait GWAS analysis is especially suited for complex correlated traits such as grain yield and its components. Hence, the combination of molecular biology and crop physiology disciplines with traditional plant breeding, could allow to push the yield grain forward (Slafer and Araus, 2005).

The aim of this article was to use a multi-trait multi-environment GWAS analysis for agronomic and physiological yield-related traits in five environments. Specifically, we intended to: (1) tackle the understanding of a complex trait such as yield by its components and related-traits, and (2) integrate information of trait components in the GWAS analysis (multi-trait and multi-environment GWAS analysis).

3.3. MATERIALS AND METHODS

3.3.1. Germplasm and phenotypic data

A set of 384 advanced inbred lines of wheat was used in this study. A total of 186 genotypes from INIA-Uruguay, 55 genotypes from INIA-Chile and 143 genotypes from CIMMYT were used. The CIMMYT genotypes share common ancestors with the INIA-Chile genotypes.

Phenotypic evaluation of all lines was conducted in five Mediterranean environments in central Chile: Cauquenes (35º 589’ S; 72º 179’ W) under rainfed conditions in 2012; and a Santa Rosa (36º 329’ S, 71º 559’ W; 217 m.a.s.l.) under two levels of water supply (mild water stress and fully irrigated) in 2011 and 2012. The two levels of water supply for Santa Rosa-Chile were: one irrigation at tillering for the mild water stress trial and four irrigations (at tillering, flag leaf emergence, heading, and middle grain filling) of 50 mm each for the fully irrigated trials. Annual precipitation in Cauquenes was 580 and 600 mm in 2011 and 2012, respectively; and in Santa Rosa was 736 and 806 mm, in 2011 and 2012, respectively (Mora et al., 2015).

The experimental design used in all trials was an alpha-lattice design with two replications and 20 incomplete blocks with 20 genotypes each. The plot size consisted of five rows of 2 m long and 0.2 m distance among rows, and the sowing rate was of 400 plants per square meter. The experiments were kept free of weeds and diseases (see Lado et al., 2013 for details of the environments). The traits evaluated were (Table 1): chlorophyll content (SP) using a SPAD 502 (Minolta Spectrum Technologics Inc., Plainfield, IL, USA) at different dates (first at anthesis, and then once a week while the plant still presented green leaves, between 80 and 130 days after sowing), plant height (PH) from the base of the plant to the spike insertion (cm) between Zadoks 8.5 and Zadoks 9.1, days to heading (DH) was recorded when 50% of culms showed emerged ears (days), photosynthetically active radiation intercepted (PAR) measured around flowering time (Zadoks 5.9 – 6.0, %) using AccuPAR Ceptometer Model LP-80 (Decagon Devices, Inc., WA, USA), leaf area index (LAI) measured around flowering time (Zadoks 5.9 – 6.0) using AccuPAR Ceptometer Model LP-80 (Decagon Devices, Inc., WA, USA), specific leaf area (SLA, $cm^2\ g^{-1}$) measured at flag leaf emergency as the relationship between leaf area and leaf dry weight, grain yield (YLD, $g\ m^{-2}$), grains per spike (GS, number) determined from the kernels harvested and weighted for the 25 spikes randomly chosen from each experimental unit, thousands kernel weight (TKW, g) determined from 25 spikes randomly chosen from each experimental unit, spikes per square meter (SPM, number) between Zadoks 8.5 and Zadoks 9.1, plants per square meter (PLM, number) before tillering, test weight (TW, g), stem weight at anthesis (ASW, g) and at maturity (MSW, g), and stem length at anthesis (ASL, cm) and at maturity (MSL, cm).

Table 1: Trait measured in each of the five environments

Trait	Abbreviation	Env. Evaluated
		SR2011D, SR2011I, SR2012D, SR2012I, C2012D
SPAD	SP	
Height (cm)	PH	
Days to heading (days)	DH	
Photosynthetically Active Radiation intercepted (%)	PAR	
Leaf area index	LAI	
Specific leaf area (cm^2 g^{-1})	SLA	
Grain yield (g m^{-2})	YLD	
Grains per spike (number)	GS	
Thousands weight grain (g)	TKW	
Spikes per square meter (number)	SPM	
Plants per square meter (number)	PLM	
Test weight (g)	TW	
Anthesis stem weight (g)	ASW	
Maturity stem weight (g)	MSW	
Anthesis stem length (cm)	ASL	
Maturity stem length (cm)	MSL	

Environment abbreviations are Santa Rosa mild water stress and fully irrigated in 2011 (SR2011D and SR2011I respectively), Santa Rosa mild water stress and fully irrigated in 2012 (SR2012D and SR2012I respectively) and Cauquenes 2012 under rainfed conditions (C2102D). SP trait was measured in two or three different dates according to the environment (see details in Materials and Methods).

3.3.2. Genotypic data

DNA was extracted by the DNeasy Plant Maxi Kit (QIAGEN). Library construction was conducted in Kansas State University (Manhattan, Kansas) and Université Laval, Quebec, Canada using a PstI-MspI GBS protocol (Poland et al., 2012b). The sequencing was performed on an Illumina Hi-Sequation 2000 at the DNA core facility at the University of Missouri, Columbia, Missouri, and the McGill Univesity-Génome Quebec Innovation Centre (Montreal, Canada) for each set of libraries. SNPs (Single-Nucleotide Polymorphism) were obtained using the Tassel-

GBS Pipeline (Glaubitz et al., 2014). The base quality and distribution of sequences was studied with the Galaxy (http://galaxy.psu. edu/) software. SNPs with less than 50 % coverage and with minor allele frequency (MAF) lower than 2.5% were excluded. Sequences were blasted to the SyntheticxOpata map (synop) (Poland et al., 2012a) with the *blastn* function from *NCBI-BLAST+* package using the number of descriptions (restricts the number of matching descriptions) and the number of threads (threads in the server) set to one and percent of identity at 95%. A final matrix set of 28,217 SNPs was obtained.

3.3.3. Statistical analysis

3.3.3.1. Descriptive statistics, genetic correlations and heritabilities for each environment

We performed a two-step procedure for analyzing the data. First, we obtained the best linear unbiased estimators (BLUEs) for the genotypes. Second, with the adjusted means we performed the GWAS analysis. For obtaining the BLUEs, we used the following model for each trait in each environment:

[1] $$y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{k(j)} + \varepsilon_{ijk}$$

where y_{ijk} is the value for the phenotypic trait corresponding to the *i-th* genotype, *j-th* replication, and *k-th* incomplete block, μ is the overall mean, α_i is the effect of the *i-th* genotype, β_j is the effect of the *j-th* replication, $\delta_{k(j)}$ is the random effect of the *k-th* incomplete block within the *j-th* replication with $\delta_{k(j)} \sim N(0, \sigma^2_s)$, ε_{ijk} is the experimental error corresponding to the *i-th* genotype, *j-th* replication and *k-th* incomplete block with $\varepsilon_{ijk} \sim N(0, \sigma^2_e)$. Genotypic means were estimated with the function *lmer* (*lme4* package) in R statistical software (R Development Core Team, 2014). Broad sense heritabilities were estimated in R statistical software using model [1] but with genotypes as random effect (R Development Core Team, 2014). Genotypic correlations among environments were estimated with the Pearson coefficient correlation. Additionally, for a graphical representation of genotypes and traits a principal component analysis (PCA) for the phenotypic data using function *prcomp* from the *stats* package in R statistical software (R Development Core Team, 2014) was conducted, where a biplot was performed to graphically display

correlations among traits. An angle smaller than 90° among traits vectors shows positive correlation. An angle of 90° among traits vectors shows no correlation. An angle between 90° and 180° among traits vectors shows negative correlation.

3.3.3.2. Multi-trait GWAS analysis

The multi-trait GWAS analysis was conducted following Malosetti et al. (2007a) for population control and Malosetti et al. (2007b) for modeling the correlations across traits. First, GWAS analysis at each marker at a time was performed including a PCA for the genotypic data as a covariate in order to control for population structure (Malosetti et al. 2007a) and considering QTL by trait interaction (QTI). Second, all significant trait-specific QTL were subsequently included in a multi-QTL model and evaluated for main and QTI effects by stepwise selection, dropping markers with non-significant QTI. This model was implemented in R (R Development Core Team, 2014). We scaled some traits in order to keep the range of the different traits within similar magnitudes to avoid having a single trait dominating the results just because of its units. The traits we scaled were: grain yield, leaf area index, plants per square meter, anthesis stem weight, maturity stem weight, anthesis stem length and maturity stem length. We split the phenotypic dataset based on categories of traits in three subgroups for the multi-trait GWAS analysis. The first subgroup consists on grain yield components: grain yield, grains per spike, thousands weight grain, test weight, plants per squared meter and spikes per squared meter. The second consists on leaf related traits: PAR, leaf area index, specific leaf area and SPAD. The third consists on morphology and phenology traits: anthesis stem weight, maturity stem weight, anthesis stem length, maturity stem length, plant height and days to heading. Briefly, multi-trait GWAS analyses within each subgroup used PCA as a random component and modeled the correlations among traits with the following model:

[2] $$y = X\beta + Z\upsilon + e$$

where y is the phenotypic vector with individuals times traits rows and one column [nt]1, X is the fixed effects matrix, including the mean vector of each trait and SNP marker scores matrix for each trait with individuals times traits rows and a SNP

evaluated one at a time [nt,nt], β is the unknown vector of allelic effects to be estimated with traits rows and one column [nt]1, Z is the incidence matrix with individuals times traits rows and individuals times traits columns [nt,nt], υ is the vector of random background polygenic effects with individuals times traits rows and one column [nt]1, $\upsilon \sim N(0, G_0 \otimes G_1)$, where G_0 is PP' being P the loadings of the first ten significant principal components of the genotypic PCA analysis to correct for genotypic structure as Malosetti et al. (2007a); G_1 is the correlations between traits for modeling the correlations across traits as Malosetti et al. (2007b), and e is the residual error, $e \sim N(0, I\ ^2_e)$.

A False Discovery Rate (FDR) correction for multiple comparisons with a significance level of 0.05 was used (Benjamini and Hochberg 1995). The analysis was performed in R statistical software (R Development Core Team, 2014) with package *lme4*.

3.3.3.3. Multi-environment GWAS analysis

A multi-environment approach was used to understand water-stress by genotype interaction. Due to unbalanced data, only a subset of the traits were used for this study: YLD, GS, TKW, PH, TW and SPM. The multi-environment GWAS analysis was also conducted following Malosetti et al. (2007a; b) and evaluating first the QTL by environment interaction (QEI) and then including environment-specific QTL in a multi-QTL model for main and interaction effects similar to model [2] but with a slight modification:

[3] $$y = X\beta + Z\upsilon + e$$

where y is the phenotypic vector with individuals times environments rows and one column [ne]1, X is the fixed effects matrix with the mean vector of each environment and SNP marker scores matrix for each environment with individuals times environments rows and a SNP evaluated one at a time [ne,ne], β is the unknown vector of allelic effects to be estimated with environments rows and one column [ne]1, Z is the incidence matrix with individuals times environments rows and individuals times environments columns [ne,ne], υ is the vector of random background polygenic effects with individuals times environments rows and one

column [ne]1, $\upsilon \sim N(0, G_0 \otimes G_1)$, where G_0 is PP' being P the loadings of the first ten significant principal components of the genotypic PCA analysis to correct for genotypic structure as Malosetti et al. (2007a); G_1 is the correlations between environments for modeling the correlations across environments as Malosetti et al. (2007b), and e is the residual error, $e \sim N(0, I \sigma^2_e)$.

A False Discovery Rate (FDR) correction for multiple comparisons with a significance level of 0.05 was used (Benjamini and Hochberg 1995). The analysis was performed in R statistical software (R Development Core Team, 2014) with package *lme4*.

3.4. RESULTS

3.4.1. Genetic correlations and heritabilities

Broad sense heritabilities were relatively high for most traits and similar across environments (Table 2). However, SLA had low heritabilities in the two environments evaluated (under mild water-stress and fully irrigated at Santa Rosa in 2011), MSW had low heritabilities under irrigated conditions and YLD had lower heritabilities in Cauquenes and Santa Rosa 2011.

Correlations patterns evidenced in the biplots were more similar (with some exceptions) between trails of the same year (i.e. C2012D and SR2012I) and less similar between trails with the same irrigation treatments (i.e. SR2012I and SR2011I, Figure1, Figure 2, Supplementary Figure 1, Supplementary Figure 2, Supplementary Figure 3). Genotypes were not grouped by their origin showing admixture (Figure 1, Figure 2, Supplementary Figure 1, Supplementary Figure 2, Supplementary Figure 3).

In general, correlations between traits among Santa Rosa environments were larger than with Cauquenes environment (data not shown).

We included in the biplots two extra traits calculated from the traits measured, in order to evaluate a better picture of the grain yield genetic basis: grains per square meter (GPM) as the product of spikes per square meter (SPM) and grains per spike (GS), and post-anthesis stem weight (PASW) as the difference between maturity stem weight (MSW) and anthesis stem weight (ASW).

Table 2: Broad sense heritability (H^2) for all traits in each environment

Trait	SR2011D	SR2011I	SR2012D	SR2012I	C2012D
SP1	0.78	0.66	0.66	0.71	0.71
SP2	0.29	0.63	0.60	0.70	0.49
SP3	-	0.18	0.49	0.54	-
PH	0.75	0.78	0.88	0.77	0.46
DH	0.96	0.97	0.91	0.93	-
PAR	0.44	0.45	0.41	0.49	-
LAI	0.30	0.41	0.66	0.63	-
SLA	0.21	0.15	-	-	-
YLD	0.36	0.47	0.66	0.63	0.40
GS	0.82	0.84	-	0.75	0.48
TKW	0.84	0.93	-	0.94	0.84
SPM	0.82	0.76	0.66	0.79	0.53
PLM	0.65	0.67	-	-	-
TW	0.58	0.92	-	0.93	0.79
ASW	0.38	0.30	0.77	0.78	0.65
MSW	0.37	0.21	0.79	0.11	0.70
ASL	0.56	0.76	0.66	0.78	0.47
MSL	0.56	0.51	0.83	0.38	0.53

Environment abbreviations are Santa Rosa mild water stress and fully irrigated in 2011 (SR2011D and SR2011I respectively), Santa Rosa mild water stress and fully irrigated in 2012 (SR2012D and SR2012I respectively) and Cauquenes 2012 under rainfed conditions (C2102D). Traits abbreviations are shown in Table 1.

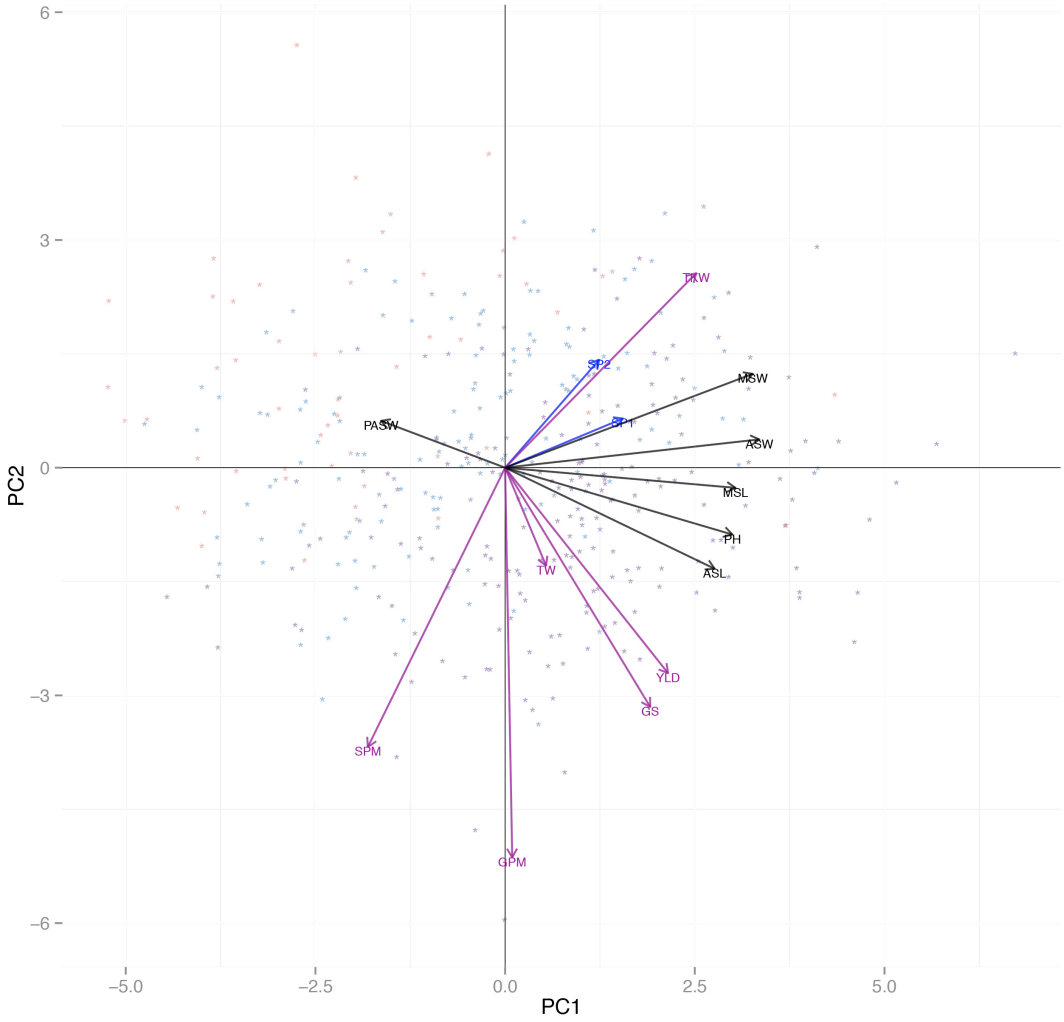


Figure 1. Biplot for the first two principal components for BLUEs for each trait in Cauquenes 2012 under rainfed conditions. Numbers indicate genotypes and its colors indicate the breeding program that each genotype belongs. Arrows indicate the traits and its colors indicate the group each trait belongs to: grain yield components (violet), leaf related traits (blue) and morphology and phenology traits (black).

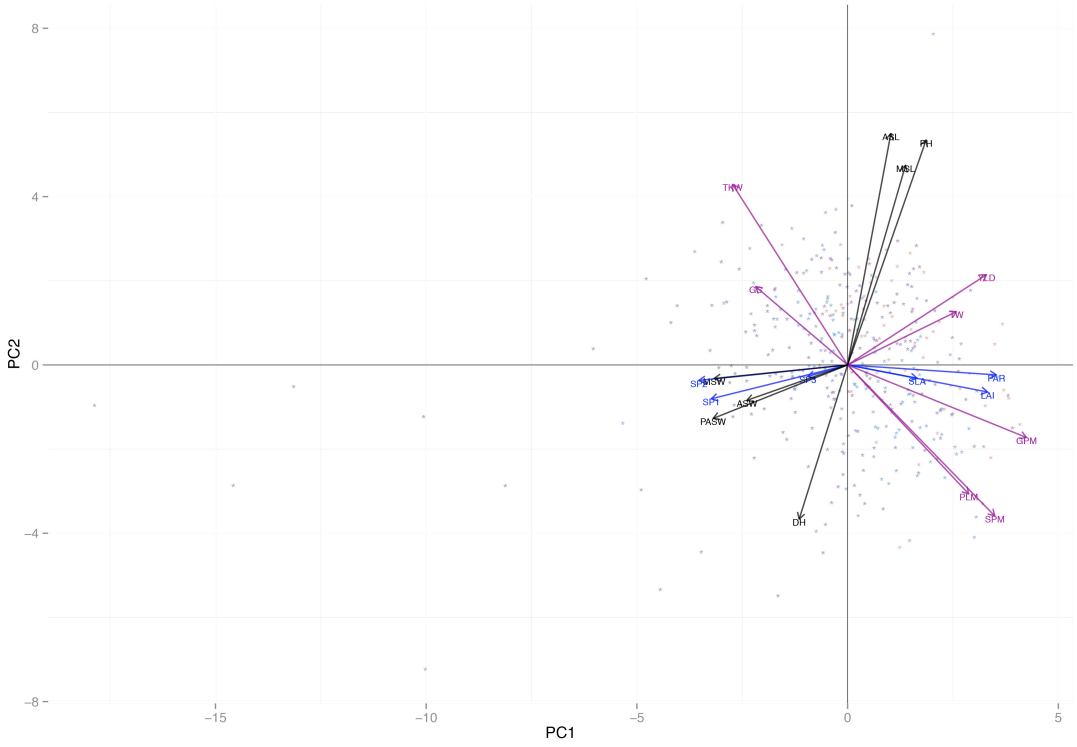


Figure 2. Biplot for the first two principal components for BLUEs for each trait in Santa Rosa 2011 fully irrigated. Numbers indicate genotypes and its colors indicate the breeding program that each genotype belongs. Arrows indicate the traits and its colors indicate the group each trait belongs to: grain yield components (violet), leaf related traits (blue) and morphology and phenology traits (black).

3.4.2. Multi-trait GWAS analysis

A total of 25 QTL (understanding for the same QTL all marker-trait associations in the same position of the same chromosome for the same group of traits) were found for all environments in all traits in the GWAS multi-trait approach, most of them in chromosome 7D (Figure 3). For grain yield components, we also found QTL in chromosome 6D bin 24 and in chromosome 7A bin 174, under mild water-stress at Santa Rosa in 2012 and 2011, respectively. For leaf related traits, we also detected a QTL in chromosome 7B bin 166 in Santa Rosa 2011 fully irrigated. Finally, for morphological and phenological traits we detected a QTL in chromosome 1B bin 105 under full irrigation in Santa Rosa 2012, and two QTLs in

chromosome 7B in 147 and 187 bins under mild water-stress in Santa Rosa 2011. Most of the QTLs were main effect QTL, with the exception of two QTL for morphology and phenology traits under mild water-stress in Santa Rosa 2011 (Figure 4). Those QTL presented a positive effect for some traits and a negative effect for others. An example of one of those two QTL is the associated with the marker iniaGBS78612 in chromosome 7B bin 147, which had a positive effect for PH, ASW, ASL, MSW and MSL but a negative effect for DH.

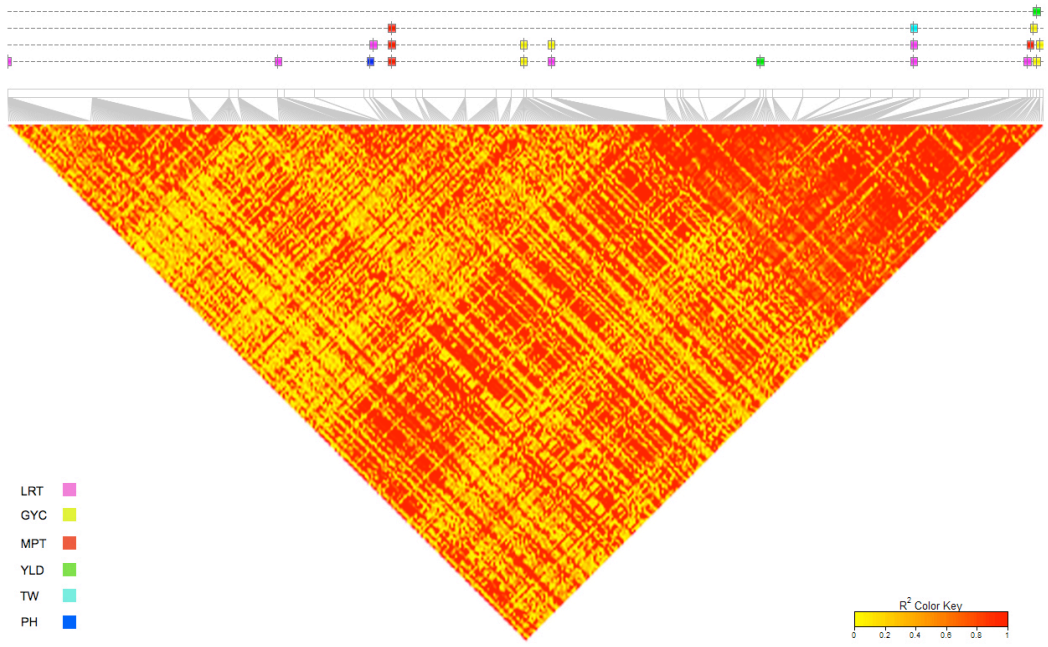


Figure 3. QTL detected in chromosome 7D and linkage disequilibrium (LD) between markers in that chromosome. QTL were detected for groups of traits in a multi-trait GWAS analysis: leaf related traits (LRT, pink), grain yield components (GYC, yellow), morphology and phenology traits (MPT, red); and for traits in a multi-environment GWAS analysis: grain yield (YLD, green), test weight (TW, sky-blue) and plant height (PH, blue).

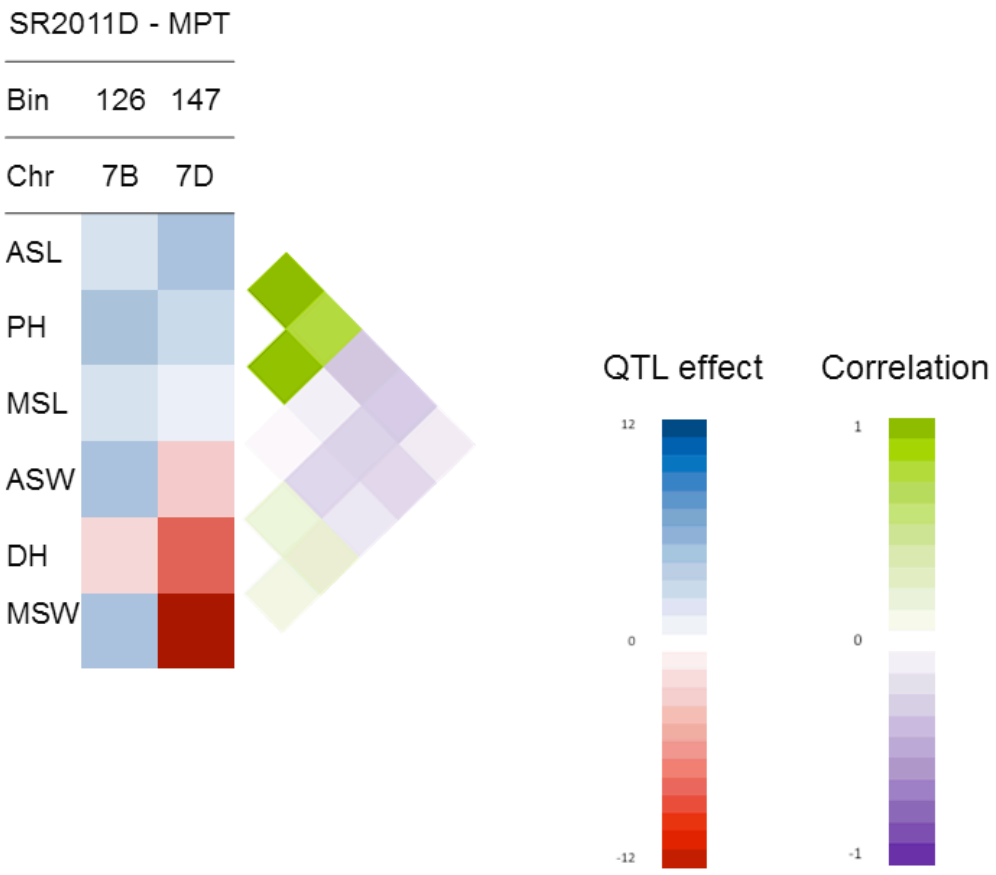


Figure 4. GWAS multi-trait profile plot for morphology and phenology traits (MPT) in Santa Rosa 2011 mild water-stress (SR2011D) and correlations between the traits for that environment. Columns are significant SNPs and rows are traits (see traits abbreviations in Materials and Methods). QTL effect: positive (blue), no (white) and negative (red), and correlation between traits: positive (green), no (white) and negative (violet).

3.4.3. Multi-environment GWAS analysis

QTL were detected only for some of the traits used for this study: YLD, GS, PH and TW. One QTL was found for PH, GS and for TW, and two QTL for YLD. Most QTL were found at chromosome 7D (Figure 3), except for GS that we detected a QTL in chromosome 6B bin 202. Only main effect QTL were found for each trait. Some QTL detected for these traits were in the same bin as the QTL detected with

the multi-trait approach. An example is the QTL detected for TW in chromosome 7D bin 296 that is in the same position of two QTL detected in LRT, one in Santa Rosa 2011 mild water-stress and the other in Cauquenes 2012 under rainfed conditions (Figure 3).

3.5. DISCUSSION

3.5.1. Genetic correlations and heritabilities

Considering grain yield components, several studies reported that the number of grains per unit of area is what mostly determines the yield performance in cereals like wheat and barley (Abeledo et al., 2003; Bustos et al., 2013; de San Celedonio et al., 2014). Nevertheless, we found a positive and higher correlation between GS and YLD than between GPM and YLD, thus number of grains per spike is determinant of the yield performance of the wheat panel evaluated in this work. Additionally, SPM was negative correlated with TKW in all environments, probably due to a compensation mechanism (Figure 1, Figure 2, Supplementary Figure 1, Supplementary Figure 2, Supplementary Figure 3). Therefore, we further analyzed the relationship between yield components (Figure 5), highlighting top 10% genotypes for YLD (38 genotypes). We found negative correlation between TKW and GPM for all environments (Figure 5). Then, analyzing if the compensation mechanisms were between TKW and GS or TKW and SPM, we detected that the negative correlation for all environments was between TKW and SPM, coincident with correlations between those traits presented in the biplots (Figure 1, Figure 2, Supplementary Figure 1, Supplementary Figure 2, Supplementary Figure 3). That competition could be the consequence of source limited conditions, leading to a less number of spikes per area (Bustos et al., 2013). Nevertheless, although the top 10% genotypes presented the compensation mechanism, they performed better in terms of grain yield because of larger TKW values or larger GMP values (Figure 5).

Considering Cauquenes 2012 under rainfed conditions (Figure 1), we observed that SPAD, which is related to chlorophyll content (Bannari et al., 2007), was closely associated to TKW, and also to ASW and MSW. Higher TKW could be a consequence of a delay in the senescence because light interception by green leaf

area is the source of carbohydrates for grain filling, which are reduce due to leaf senescence (Moschen et al., 2014). The importance of TKW as a grain yield component in this environment could be a consequence of the low tillering survival (Figure 5).

In fully irrigated conditions, PASW should be more similar to ASW than in rainfed conditions, because carbon assimilation is not affected by stress and therefore storage in stems is not reduced (Blum, 1998). Considering Santa Rosa 2011 fully irrigated (Figure 1), we found that SPAD was closely associated to ASW and MSW, but additionally to PASW, and thus we believed that the closest relationship between ASW and PASW is due to the not-stress growing conditions.

An increase in LAI is expected to be followed by an increase in PAR, until the 90% of the photosynthetically active radiation is reached and thus no improving in PAR is expected (Hipps, 1983). As we found positive and strong correlation between those traits in Santa Rosa 2011 fully irrigated (Figure 2), we believe that the 90% of the photosynthetically active radiation was not reached at the time the traits were measured (Zadoks 5.9 – 6.0).

Higher correlations between leaf related traits and grain yield is expected in fully irrigated environments because no water stress is affecting the source, and therefore an increase in the light intercepted is followed by an increase in grain yield (Aparicio et al., 1999). We therefore found that YLD and TW were positive and strong correlated with PAR, LAI and SLA in Santa Rosa 2011 fully irrigated (Figure 2).

Breeding for shortened cycles in Mediterranean conditions has been proved to be a successful strategy (Araus et al., 2002). Therefore, the negative correlation between DH and YLD should be an index of higher yields in Mediterranean environments. We found for Santa Rosa 2011 fully irrigated a negative correlation between DH and YLD and low correlations for Santa Rosa mild water stress for 2011 and 2012.

Strong and positive correlations between ASL, MSL and PH could be explained by a not longer growth of the stem after anthesis in crops like wheat.

Studying the relationship between sink and source for the top 10% genotypes in each environment in order to evaluate if there were differences between the environments, we considered the ratio between GS and ASW for the environments where the traits were measured (Santa Rosa mild water stress and fully irrigated 2011, Santa Rosa fully irrigated 2012 and Cauquenes 2012). Stem weight at anthesis (ASW, related to water soluble carbohydrate content) is a source for grain growth, because it is supported by remobilization of carbohydrate reserve from the stem and photosynthesis of leaves and spike (del Pozo et al., 2012). We found higher ratios for Santa Rosa 2011 and 2012 fully irrigated (31.4 and 32.3 respectively), and lower ratios for Santa Rosa 2011 mild water stress and Cauquenes 2012 (27.6 and 19.3 respectively). Therefore, we analyzed separately GS and ASW for those environments and we found that Cauquenes 2012 was the environment with the lower median for GS but higher for ASW (data not shown). Consequently, these results could indicate that in water stress environments (Santa Rosa 2011 mild water stress and Cauquenes 2012), as tillering survival are affected (Figure 5), higher ASW is obtained. These results combined with the results found from del Pozo et al., (2012) in barley evaluated in the same environments, could indicated that higher ASW is found under water-stress conditions due to higher carbohydrates in the stem and lower mobilization of carbohydrates to the grains.

We used two contrasting Mediterranean-type environments in this study (Cauquenes vs. Santa Rosa). In Cauquenes, plants face terminal drought stress, then early flowering (shortening life cycle) is an escape mechanism from dehydration (del Pozo et al., 2012). We found larger correlations between fully irrigated and mild water stress environments in Santa Rosa, and smaller correlations with Cauquenes for most of the traits (data not shown). The lowest correlations between Cauquenes and Santa Rosa environments could be the consequence of being different Mediterranean-type environments. Additionally, the differences between these locations were evidenced in Cauquenes, which presented less SPM, because drought conditions in Cauquenes reduced tillering survival and therefore SPM being the compensation mechanism between grain yield components more clear (Figure 5).

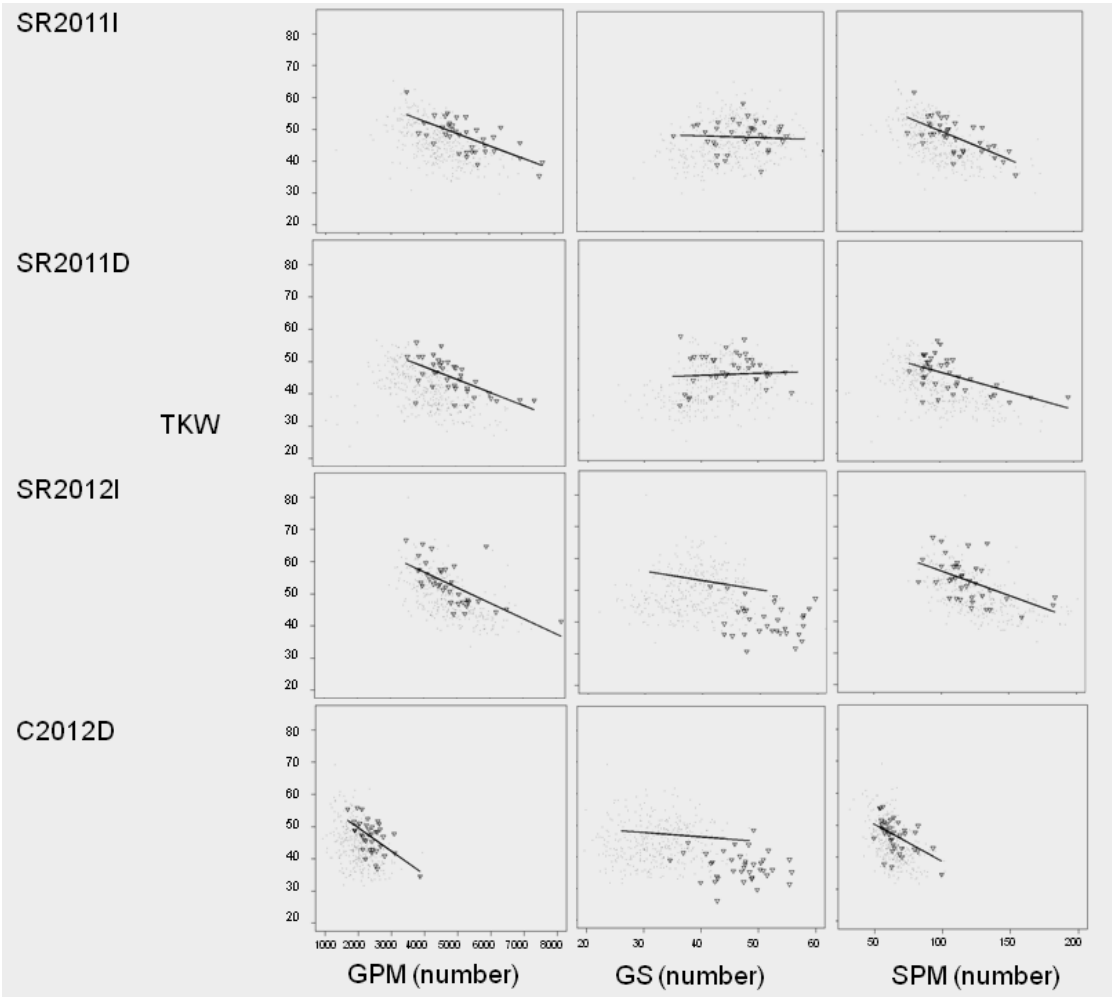


Figure 5. Relationship among gran yield (YLD) components: TKW (thousands kernel weight), grains per spike (GS), spikes per square meter (SPM), and grains per square meter (GPM) as the product between GS and SPM in each environment. The 10% genotypes for YLD are highlighted in each environment. Environments are: Santa Rosa 2011 fully irrigated (SR2011I), Santa Rosa 2011 mild water stress (SR2011D), Santa Rosa 2012 fully irrigated (SR2012I) and Cauquenes 2012 under rainfed conditions (C2012D).

3.5.2. Multi-trait GWAS analysis

QTL for physiological traits that were pleiotropic or closely link or that showed interactions were detected (Figure 3). The strategy of multi-trait for GWAS analysis

allowed us to identify mostly main effect QTL for physiological, correlated traits. It was expected that positive and strong correlated traits had main effect QTL (Alimi et al. 2013). Although within the subgroups not all traits were strong and positive correlated, main effect QTL found could indicate that the development of the crop is mostly affected by the same genetics basis for biology related traits. The sum of all these QTL and some QTI, it is what finally produces wheat grain yield.

3.5.3. Multi-environment GWAS analysis

Only main effects QTL were detected for the traits evaluated among the environments (Figure 3). This suggests that although grain yield is a complex trait compounded by other traits and affected by different biotic and abiotic factors (Kjaer and Jensen, 1996; Singh et al., 2008; Mathews et al., 2008; van Eeuwijk et al., 2010; Alimi et al., 2012, 2013; Malosetti et al., 2013), no QEI were detected and therefore introducing those QTL in the breeding program's populations and evaluating the lines in one environment should improve grain yield performance for the lines selected in all environments evaluated. Although we expected QTL by environment interaction in this analysis because of the different water conditions for the crop, the rainfall in spring occurred in Santa Rosa 2012, resulted in Santa Rosa 2012 mild water stress being similar to Santa Rosa 2012 fully irrigated, and thus those environments were not as contrasting as we expected for that year.

3.5.4. Previous QTL reported

Since most QTL in this analysis were found in chromosome 7D, we focused on finding previous QTL reported in this chromosome. Several studies found QTL for yield and yield-related traits on 7D. QTL for grain weight, spikes number, fertile spikelet number per spike, sterile spikelet number per spike were found in 7D associated with markers Xgdm67, Xgwm428 and Xwmc31 (Li et al., 2007). Additionally, QTL for grain weight was found between Xgwm295 and Xgwm1002 in chromosome 7D (Röder et al., 2008). For biomass, single kernel weight and test weight QTL were found in chromosome 7D between positions 170 and 172.2 cM based in marker physical map (http://www.cerealdb.uk.net/, Edae et al., 2014). In

other study, QTL for grain width was detected in chromosome 7D associated with Xgwm295 and Xgwm437; for grain volume associated with Xgdm86c and Xgwm885 and grain vertical perimeter associated with Xgdm86c and Xgwm885 (Tyagi et al., 2014).

Additionally, we compared the QTL found in this analysis with the QTL found in Mora et al., (2015), that analyzed the same population evaluated in the same environments considering grain yield components and carbon isotope discrimination. The QTL found for morphology and phenology traits with QTL by trait interaction in chromosome 7B bin 126, was also detected by Mora et al., (2015) for kernels per spike for Santa Rosa 2011 and 2012 and for the same trait when they performed a multi-environment analysis considering all environments. The QTL found for grain yield components in chromosome 7A bin 174 was closely to a QTL detected by Mora et al., (2015) for thousand kernel weight. Finally, a QTL found for leaf related traits in chromosome 7D bin 0, was also detected by Mora et al., (2015) for thousand kernel weight. Although Mora et al., (2015) detected more QTL by environment interactions that we did, they did no modeled the correlation between environments, assuming a diagonal covariance matrix, and therefore more interactions were found.

3.6. CONCLUSIONS

In conclusion, primarily main effect QTL were found with a multi-trait GWAS analysis by studying correlated traits within subgroups (grain yield components, leaf related traits, and morphology and phenology traits). Those main effects QTL found could indicate that the development of the crop is mostly affected by the same genetics basis for biology related traits. The same compensation mechanisms between grain yield components were detected, where more SPM reduced the values of TKW, being more evident in Cauquenes 2012 dry-agricultural area. The multi-environment approach allowed us to detect main effect QTL among the environments, implying that QTL with positive effect for grain yield in one environment, could be introduced in the breeding program's populations and improve the genotypes performance for all environments evaluated. Further analysis on QTL regions should be perform for a better resolution of the QTL found and a

later introduction of these QTL in the breeding programs involved in order to improve grain yield in Mediterranean environments.

3.7. REFERENCES

Abeledo, L.G., D.F. Calderini, and G.A. Slafer. 2003. Genetic improvement of barley yield potential and its physiological determinants in Argentina (1944–1998). Euphytica 130(3):325–334.

Acreche, M.M., G. Briceño-Félix, J.A. Martín Sánchez, and G.A. Slafer. 2008. Radiation interception and use efficiency as affected by breeding in Mediterranean wheat. Field Crop Res. 110:91–97.

Alimi, N.A., M.C.A.M. Bink, J.A. Dieleman, J.J. Magán, A.M. Wubs, A. Palloix, and F.A. van Eeuwijk. 2013. Multi-trait and multi-environment QTL analyses of yield and a set of physiological traits in pepper. Theor. Appl. Genet. 126(10):2597–625.

Alimi, N.A., M.C.A.M. Bink, J.A. Dieleman, M. Nicolaï, M. Wubs, E. Heuvelink, J. Magan, R.E. Voorrips, J. Jansen, P.C. Rodrigues, G.W.A.M. Heijden, a. Vercauteren, M. Vuylsteke, Y. Song, C. Glasbey, A. Barocsi, V. Lefebvre, A. Palloix, and F.A. van Eeuwijk. 2012. Genetic and QTL analyses of yield and a set of physiological traits in pepper. Euphytica 190(2):181–201.

Aparicio, N., D. Villegas, J. Casadesus, J.L. Araus, and C. Royo. 1999. Spectral Vegetation Indices as Nondestructive Tools for Determining Durum Wheat Yield. Agron. J. 91(202000):83–91.

Araus, J.L., G.A. Slafer, M.P. Reynolds, and C. Royo. 2002. Plant breeding and drought in C3 cereals: What should we breed for? Ann. Bot. 89(SPEC. ISS.):925–940.

Bannari, A., K.S. Khurshid, K. Staenz, and J.W. Schwarz. 2007. A comparison of hyperspectral chlorophyll indices for wheat crop chlorophyll content estimation using laboratory reflectance measurements. IEEE Trans. Geosci. Remote Sens. 45(10):3063–3074.

Barnabás, B., K. Jäger, and A. Fehér. 2008. The effect of drought and heat stress on reproductive processes in cereals. Plant. Cell Environ. 31(1):11–38.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc. 57(1):289–300.

Blum, A. 1998. Improving wheat grain filling under stress by stem reserve mobilisation. Euphytica 100(1):77–83.

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(3):1801–13.

Bustos, D.V., A.K. Hasan, M.P. Reynolds, and D.F. Calderini. 2013. Combining high grain number and weight through a DH-population to improve grain yield potential of wheat in high-yielding environments. F. Crop. Res. 145:106–115.

Chen, X.M. 2005. Epidemiology and control of stripe rust [Puccinia striiformis f. sp. tritici] on wheat. Can. J. Plant Pathol. 27:314–337.

Del Pozo, A., D. Castillo, L. Inostroza, I. Matus, A.M. Méndez, and R. Morcuende. 2012. Physiological and yield responses of recombinant chromosome substitution lines of barley to terminal drought in a Mediterranean-type environment. Ann. Appl. Biol. 160(2):157–167.

De San Celedonio, R.P., L.G. Abeledo, and D.J. Miralles. 2014. Identifying the critical period for waterlogging on yield and its components in wheat and barley. Plant Soil 378(1-2):265–277.

Edae, E.A., P.F. Byrne, S.D. Haley, M.S. Lopes, and M.P. Reynolds. 2014. Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. Theor. Appl. Genet. 127(4):791–807.

El-Soda, M., W. Kruijer, M. Malosetti, M. Koornneef, and M.G.M. Aarts. 2014. Quantitative trait loci and candidate genes underlying genotype by environment interaction in the response of Arabidopsis thaliana to drought. Plant, Cell & Environment, 38(3):585-599.

Ewert, F., M.D.A. Rounsevell, I. Reginster, M.J. Metzger, and R. Leemans. 2005. Future scenarios of European agricultural land use. Agric. Ecosyst. Environ. 107(2-3):101–116.

FAOSTAT. 2014. Production. Available at http://faostat3.fao.org/home. (verified 03 June 2015). FAO, Rome, Italy.

Fischer, R.A. 2007. Understanding the physiological basis of yield potential in wheat. J. Agric. Sci. 145(02):99.

García, G.A., A.K. Hasan, L.E. Puhl, M.P. Reynolds, D.F. Calderini, and D.J. Miralles. 2013. Grain Yield Potential Strategies in an Elite Wheat Double-Haploid Population Grown in Contrasting Environments. Crop Sci. 53(6):2577.

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(2):e90346.

Hayes, P.M., B.H. Liu, S.J. Knapp, F. Chen, B. Jones, T. Blake, J. Franckowiak, D. Rasmusson, M. Sorrells, S.E. Ullrich, D. Wesenberg, and A. Kleinhofs. 1993. Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. Theor. Appl. Genet. 87(3):392–401.

Hipps, L. 1983. Assessing the interception of photosynthetically active ratiation in winter wheat. Agric. Meteorol. 28:253–259.

Jiang, C., and Z.B. Zeng. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140(3):1111–27.

Kjaer, B., and J. Jensen. 1996. Quantitative trait loci for grain yield and components in a cross between a six-rowed and a two-rowed barley. Euphytica. 39–48.

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 (Bethesda) 3(12):2105–14.

Li, S., J. Jia, X. Wei, X. Zhang, L. Li, H. Chen, Y. Fan, H. Sun, X. Zhao, T. Lei, Y. Xu, F. Jiang, H. Wang, and L. Li. 2007. A intervarietal genetic map and QTL analysis for yield traits in wheat. Mol. Breed. 20(2):167–178.

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(March):44.

Malosetti, M., J.M. Ribaut, M. Vargas, J. Crossa, and F.A. van Eeuwijk. 2007b. 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., C.G. van der Linden, B. Vosman, and F.A. van Eeuwijk. 2007a. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics 175(2):879–89.

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(7):1077–91.

Mora, F., D. Castillo, B. Lado, I. Matus, J. Poland, F. Belzile, J. von Zitzewitz, and A. del Pozo. 2015. Genome-wide association mapping of agronomic traits and carbon isotope discrimination in a worldwide germplasm collection of spring wheat using SNP markers. Mol. Breed. 35:69.

Moschen, S., S. Bengoa Luoni, N.B. Paniego, H.E. Hopp, G.A.A. Dosio, P. Fernandez, and R.A. Heinz. 2014. Identification of Candidate Genes Associated with Leaf Senescence in Cultivated Sunflower (Helianthus annuus L.). PLoS One 9(8):e104379.

Mueller, N.D., J.S. Gerber, M. Johnston, D.K. Ray, N. Ramankutty, and J.A. Foley. 2012. Closing yield gaps through nutrient and water management. Nature 490(7419):254–7.

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(2):e32253.

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 Genome J. 5(3):103.

R Development Core Team. 2014. R: A language and environment for statistical computing, reference index version 3.1.3. Available at http://www.R-

project.org (verifi ed 03 June 2015). R Foundation for Statistical Computing, Vienna, Austria.

Reynolds, M., J. Foulkes, R. Furbank, S. Griffiths, J. King, E. Murchie, M. Parry, G. Slafer. 2012. Achieving yield gains in wheat. Plant, Cell & Environment. 35:1799–1823.

Röder, M.S., X.Q. Huang, and A. Börner. 2008. Fine mapping of the region on wheat chromosome 7D controlling grain weight. Funct. Integr. Genomics 8(1):79–86.

Singh, R.P., D.P. Hodson, J. Huerta-espino, Y. Jin, P. Njau, R. Wanyera, S.A. Herrera-foessel, and R.W. Ward. 2008. Will Stem Rust Destroy the World's Wheat Crop? Advances in Agronomy 98:271-309.

Slafer, G.A., and J.L. Araus. Physiological traits for imrpoving wheat yield under a wide range of conditions. Scale and Complexity in Plant Systems Research: Gene-Plant-Crop Relations. 21:147–156.

Tyagi, S., R.R. Mir, H.S. Balyan, and P.K. Gupta. 2014. Interval mapping and meta-QTL analysis of grain traits in common wheat (Triticum aestivum L.). Euphytica 201(3):367–380.

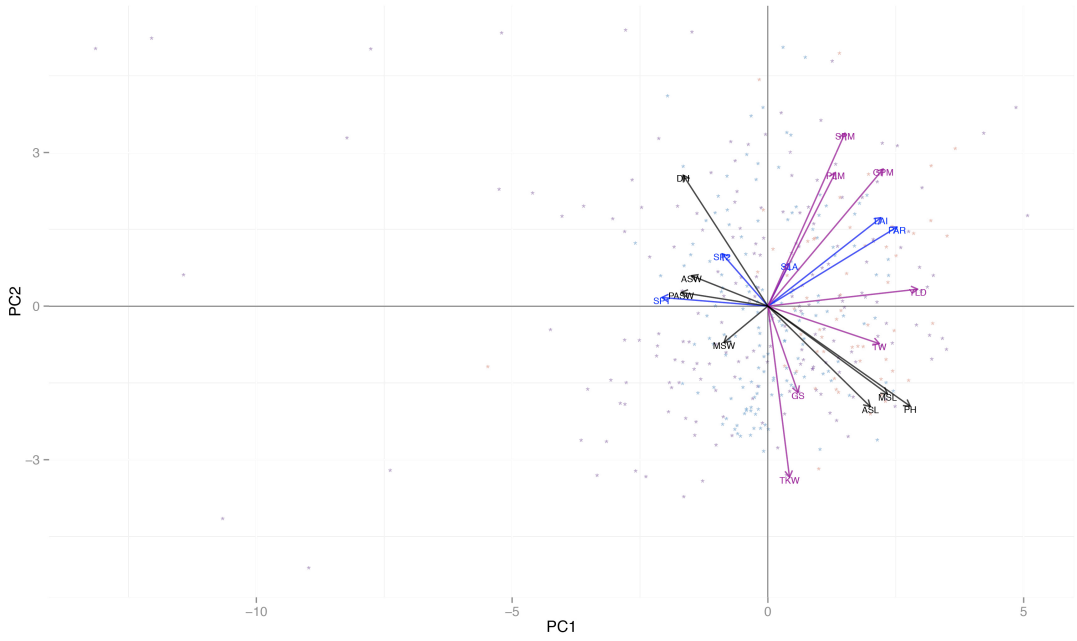
van Eeuwijk, F.A., M.C.A.M. Bink, K. Chenu, and S.C. Chapman. 2010. Detection and use of QTL for complex traits in multiple environments. Curr. Opin. Plant Biol. 13(2):193–205.

Windels, C.E. 2000. Economic and social impacts of fusarium head blight: changing farms and rural communities in the northern great plains. Phytopathology 90(1):17–21.

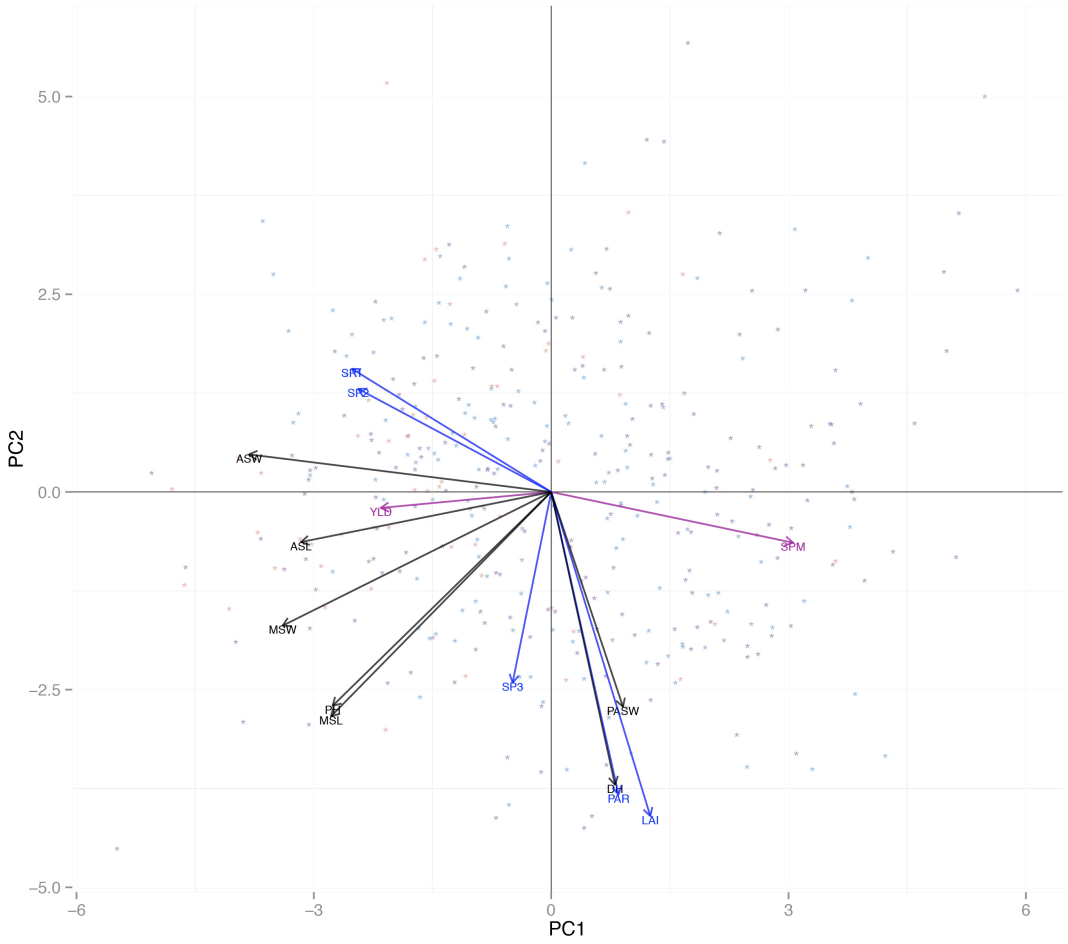
Zhu, C., M. Gore, E.S. Buckler, and J. Yu. 2008. Status and Prospects of Association Mapping in Plants. Plant Genome J. 1(1):5.

3.8. SUPPLEMENTARY FIGURES

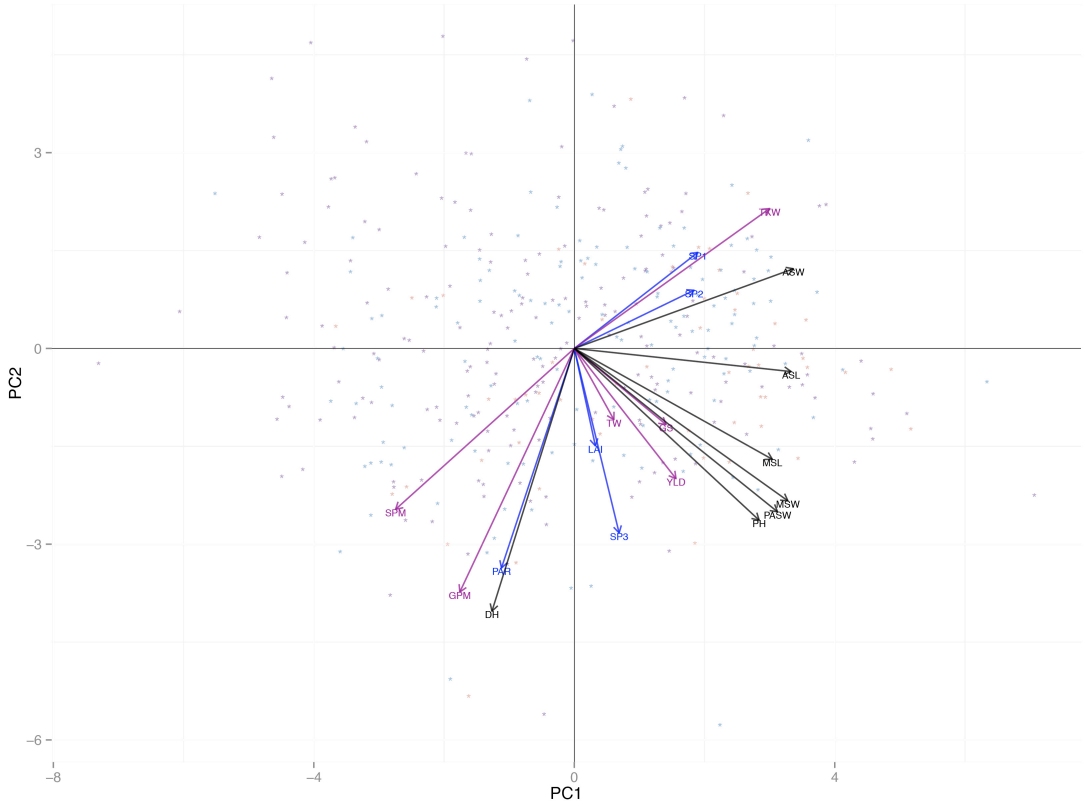
3.8.1. <u>Supplementary Figure 1.</u> Biplot for the first two principal components of the traits and lines for Santa Rosa 2011 mild water stress. Asterisks indicate genotypes and its colors indicate the breeding program that each genotype belongs to: coral for CIMMYT, blue for INIA-Chile and purple for INIA-Uruguay. Arrows indicate the traits and its colors indicate the group each trait belongs to: grain yield components (violet), leaf related traits (blue) and morphology and phenology traits (black).



3.8.2. <u>Supplementary Figure 2.</u> Biplot for the first two principal components of the traits and lines for and Santa Rosa 2012 mild water stress. Asterisks indicate genotypes and its colors indicate the breeding program that each genotype belongs to: coral for CIMMYT, blue for INIA-Chile and purple for INIA-Uruguay. Arrows indicate the traits and its colors indicate the group each trait belongs to: grain yield components (violet), leaf related traits (blue) and morphology and phenology traits (black).



3.8.3. <u>Supplementary Figure 3.</u> Biplot for the first two principal components of the traits and lines for Santa Rosa 2012 fully irrigated. Asterisks indicate genotypes and its colors indicate the breeding program that each genotype belongs to: coral for CIMMYT, blue for INIA-Chile and purple for INIA-Uruguay. Arrows indicate the traits and its colors indicate the group each trait belongs to: grain yield components (violet), leaf related traits (blue) and morphology and phenology traits (black).



4. DISCUSIÓN GENERAL Y CONCLUSIONES GLOBALES

Al ser el trigo uno de los cultivos de mayor relevancia mundial (FAOSTAT, 2014), varios investigadores y programas de mejoramiento están trabajando conjuntamente con el fin de aumentar su productividad y así poder sobrellevar la demanda poblacional de alimentos creciente (Mueller et al., 2012). Esta tesis se enmarca dentro del programa de mejoramiento de trigo en Uruguay, donde conjuntamente se trabajó con el Instituto Nacional de Investigación Agropecuaria (INIA), con el fin de contribuir al mejoramiento del rendimiento del trigo.

Considerando el primer objetivo, al comparar la performance de la imputación de la matriz genotípica en el mapeo asociativo cuando no hay un panel de referencia y una gran proporción de datos faltantes se presenta como en GBS, detectamos que dicha imputación puede introducir un sesgo en el análisis de mapeo. Comparando el poder de detección de QTL y la tasa de falsos positivos utilizando matrices imputadas y sin imputar (50% de datos faltantes), la performance del análisis de mapeo disminuyó cuando fue realizado con una matriz imputada. Por lo tanto, cuando no se presenta un panel de referencia y hay varios datos faltantes en la matriz genotípica, el mapeo asociativo debería realizarse sin la imputación de dicha matriz.

Con respecto al objetivo dos, sobre evaluar los factores genéticos asociados a variables agronómicas y fisiológicas en trigo considerando la interacción genotipo por ambiente, detectamos principalmente QTL de efecto principal tanto para el análisis multi-carácter como para el multi-ambiente, y algunas interacciones QTL por carácter. Los QTL de efecto principal para el análisis multi-carácter, podrían indicar que el desarrollo del cultivo es principalmente afectado por la misma base genética para variables biológicas asociadas. En relación al análisis multi-ambiente, los QTL de efecto principal detectados podrían incorporarse en las líneas de los programas de mejoramiento evaluadas para uno de los ambientes analizados, implicando que si un QTL mejoró el rendimiento en grano para ese ambiente, también lo hará para dichas líneas en los otros ambientes evaluados en este estudio. Análisis más detallados en las regiones de los QTL detectados deberían ser realizados, para obtener una mejor

resolución de los QTL y así poder introducirlos para mejorar el rendimiento de trigo en regiones Mediterráneas.

5. BIBLIOGRAFÍA

Abeledo LG, Calderini DF, Slafer GA. 2003. Genetic improvement of barley yield potential and its physiological determinants in Argentina (1944–1998). Euphytica, 130(3): 325–334.

Acreche MM, Briceño-Félix G, Sánchez JAM, Slafer GA. 2008. Physiological bases of genetic gains in Mediterranean bread wheat yield in Spain. European Journal of Agriculture, 28(28): 162-170.

Alimi NA, Bink MCAM, Dieleman JA, Magán JJ, Wubs AM, Palloix A, van Eeuwijk FA. 2013. Multi-trait and multi-environment QTL analyses of yield and a set of physiological traits in pepper. Theoretical and Applied Genetics, 126: 2597–625.

Alimi NA, Bink MCAM, Dieleman JA, Nicolaï M, Wubs M, Heuvelink E, Magan J, Voorrips RE, Jansen J, Rodrigues PC, Heijden GWAM, Vercauteren A, Vuylsteke M, Song Y, Glasbey C, Barocsi A, Lefebvre V, Palloix A, Eeuwijk FA. 2012. Genetic and QTL analyses of yield and a set of physiological traits in pepper. Euphytica, 190: 181–201.

Almeida MAA, Oliveira PSL, Pereira TV, Krieger JE, Pereira AC. 2011. An empirical evaluation of imputation accuracy for association statistics reveals increased type-I error rates in genome-wide associations. BMC Genetics, 12:10.

Aparicio N, Villegas D, Casadesus J, Araus JL, Royo C. 1999. Spectral Vegetation Indices as Nondestructive Tools for Determining Durum Wheat Yield. Agronomical Journal, 91(202000): 83–91.

Araus JL, Slafer GA, Reynolds MP, Royo C. 2002. Plant breeding and drought in C3 cereals: What should we breed for? Annuals of Botany, 89(SPEC. ISS.): 925–940.

Aulchenko YS, Struchalin MV, van Duijn CM. 2010. ProbABEL package for genome-wide association analysis of imputed data. BMC Bioinformatics, 11:134.

Bannari A, Khurshid KS, Staenz K, Schwarz JW. 2007. A comparison of hyperspectral chlorophyll indices for wheat crop chlorophyll content estimation

using laboratory reflectance measurements. IEEE Trans. Geosci. Remote Sens., 45(10): 3063–3074.

Barnabás B, Jäger K, Fehér A. 2008. The effect of drought and heat stress on reproductive processes in cereals. Plant, Cell & Environment, 31(1): 11–38.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, 57(1): 289–300.

Blum A. 1998. Improving wheat grain filling under stress by stem reserve mobilisation. Euphytica 100(1): 77–83.

Bennett D, Reynolds M, Mullan D, Izanloo A, Kuchel H, Langridge P, Schnurbusch T. 2102. Detection of two major grain yield QTL in bread wheat (Triticum aestivum L.) under heat, drought and high yield potential environments. Theoretical and Applied Genetics, 125: 1473–85.

Bernardo R. Breeding for Quantitative Traits in Plants. 2nd ed. Minnesota: Stemma Press; 2010.

Boer MP, Wright D, Feng L, Podlich DW, Luo L, Cooper M, van Eeuwijk FA. 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: 1801–13.

Browning SR. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. Human Genetics, 124(5): 439–50.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics, 81: 1084–1097.

Bustos DV, Hasan AK, Reynolds MP, Calderini DF. 2013. Combining high grain number and weight through a DH-population to improve grain yield potential of wheat in high-yielding environments. Field Crops Research, 145: 106–115.

Chao S, Dubcovsky J, Dvorák J, Luo MC, Baenziger SP, Matnyazov R, Clark DR, Talbert LE, Anderson JA, Dreisigacker S, Glover K, Chen J, Campbell K, Bruckner PL, Rudd JC, Haley S, Carver BF, Perry S, Sorrells ME, Akhunov

ED. 2010. Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (Triticum aestivum L.). BMC genomics, 11(1): 727.

Chen XM. 2005. Epidemiology and control of stripe rust [Puccinia striiformis f. sp. tritici] on wheat. Canadian Journal of Plant Pathology, 27: 314–337.

Chengsong Z, Jianming Y. 2009. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics, 182: 875–888.

Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson JT, Wanamaker S, Bozdag S, Roose ML, Moscou MJ, Chao S, Varshney RK, Sz P, Sato K, Hayes PM, Matthews DE, Kleinhofs A, Muehlbauer GJ, Deyoung J, Marshall DF, Madishetty K, Fenton RD, Condamine P, Graner A, Waugh R. 2009. Development and implementation of high-throughput SNP genotyping in barley, 13: 1–13.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Review Genetics, 12: 499–510.

De Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF. 2008. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Human Molecular Genetics, 17: 122–128.

Del Pozo A, Castillo D, Inostroza L, Matus I, Méndez AM, Morcuende R. 2012. Physiological and yield responses of recombinant chromosome substitution lines of barley to terminal drought in a Mediterranean-type environment. Annals of Applied Biology. 160(2): 157–167.

De San Celedonio, RP, Abeledo LG, Miralles DJ. 2014. Identifying the critical period for waterlogging on yield and its components in wheat and barley. Plant Soil 378(1-2): 265–277.

Dubcovsky J, Dvorak J. 2007. Genome Plasticity a Key Factor. Science 316, 1862.

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

Edae EA, Byrne PF, Haley SD, Lopes MS, Reynolds MP. 2014. Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. Theoretical and Applied Genetics, 127: 791–807.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS One, 6(5): e19379.

El-Soda M, Kruijer W, Malosetti M, Koornneef M, Aarts MGM. 2014. Quantitative trait loci and candidate genes underlying genotype by environment interaction in the response of Arabidopsis thaliana to drought. Plant, Cell & Environment, 38(3): 585-599.

Ewert F, Rounsevell MDA, Reginster I, Metzger MJ, Leemans R. 2005. Future scenarios of European agricultural land use. Agriculture, Ecosystems & Environment, 107(2-3): 101–116.

FAOSTAT. 2014. Production. [En línea]. 3 junio 2015. FAO, Rome, Italy. http://faostat3.fao.org/home.

Fischer, RA. 2007. Understanding the physiological basis of yield potential in wheat. Journal of Agricultural Science, 145(02): 99.

Flint-Garcia SA, Thornsberry JM, Buckler ES. 2003. Structure of linkage disequilibrium in plants. Annual Review of Plant Biology, 54: 357–74.

García GA, Hasan AK, Puhl LE, Reynolds MP, Calderini DF, Miralles DJ. 2013. Grain Yield Potential Strategies in an Elite Wheat Double-Haploid Population Grown in Contrasting Environments. Crop Science, 53(6): 2577–2587.

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(2): e90346.

Guan Y, Stephens M. 2008. Practical issues in imputation-based association mapping. PLoS Genetics, ,4: e1000279.

Gutiérrez L, Germán S, Pereyra S, Hayes PM, Pérez CA, Capettini F, Locatelli A, Berberian NM, Falconi EE, Estrada R, Fros D, Gonza V, Altamirano H, Huerta-Espino J, Neyra E, Orjeda G, Sandoval-Islas S, Singh R, Turkington K, Castro AJ. 2014. Multi-environment multi-QTL association mapping identifies

disease resistance QTL in barley germplasm from Latin America. Theoretical and Applied Genetics, 128:501–516.

Gutiérrez L, Cuesta-Marcos A, Castro AJ, von Zitzewitz J, Schmitt M, Hayes P. 2011. Association Mapping of Malting Quality Quantitative Trait Loci in Winter Barley: Positive Signals from Small Germplasm Arrays. The Plant Genome, 4 (3): 256-272.

Hao K, Chudin E, McElwee J, Schadt EE. 2009. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genetics, 10:27.

Hayes PM, Liu BH, Knapp SJ, et al. 1993. Quantitative trait locus effects and environmental interaction in a sample of North American barley germplasm. Theoretical and Applied Genetics, 87(3): 392-401.

He S, Zhao Y, Mette MF, Bothe R, Ebmeyer E, Sharbel TF, Reif JC, Jiang Y. 2015. Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L.). BMC Genomics, 16:1–12.

Heslot N, Rutkoski J, Poland J, Jannink J-L, Sorrells ME. 2013. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One, 8: e74612.

Hipps L. 1983. Assessing the interception of photosynthetically active ratiation in winter wheat. Agric. Meteorol., 28: 253–259.

Howie B, Marchini J, Stephens M. 2011. Genotype imputation with thousands of genomes. G3 (Bethesda), 1(6), 457–470.

Iwata H, Jannink JL. 2010. Marker Genotype Imputation in a Low-Marker-Density Panel with a High-Marker-Density Reference Panel: Accuracy Evaluation in Barley Breeding Lines. Crop Science, 50(4), 1269-1278.

Jannink JL, Iwata H, Bhat PR, Chao S, Wenzl P, Muehlbauer GJ. 2009. Marker Imputation in Barley Association Studies. The Plant Genome Journal, *2*(1): 11-22.

Jansen C, von Wettstein D, Schafer W, Kogel KH, Felk A, Maier FJ. 2005. Infection patterns in barley and wheat spikes inoculated with wild-type and trichodiene

synthase gene disrupted Fusarium graminearum. PNAS, 102(46): 16892–16897.

Jiang C, Zeng ZB. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics, 140(3): 1111–27.

Kjaer B, Jensen J. 1996. Quantitative trait loci for grain yield and components in a cross between a six-rowed and a two-rowed barley. Euphytica, 39–48.

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 (Bethesda). 3(12): 2105–2114.

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

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MACH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetics Epidemiology, 34: 816–834.

Li S, Jia J, Wei X, Zhang X, Li L, Chen H, Fan Y, Sun H, Zhao X, Lei T, Xu Y, Jiang F, Wang H, Li L. 2007. A intervarietal genetic map and QTL analysis for yield traits in wheat. Molecular Breeding, 20(2): 167–178.

Li J, Ji L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity, 95: 221–227.

Malosetti M, Ribaut J-M, van Eeuwijk F. 2013. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. Frontiers in Physiology, 4: 44.

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk F. 2007a. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics, 175(2): 879–89.

Malosetti M, Ribaut JM, Vargas M, Crossa J, van Eeuwijk F. 2007b. 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.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics, 39:906–13.

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

Mathews KL, Malosetti M, Chapman S, McIntyre L, Reynolds M, Shorter R, van Eeuwijk FA. 2008. Multi-environment QTL mixed models for drought stress adaptation in wheat. Theoretical and Applied Genetics, 117: 1077–91.

Mayer KFX, Rogers J, Dole el J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski a. J, Sourdille P, Endo TR, Kubalakova M, Ihalikova J, Dubska Z, Vrana J, Perkova R, Imkova H, Febrer M, Clissold L, McLay K, Singh K, Chhuneja P, Singh NK, Khurana J, Akhunov E, Choulet F, Alberti a., Barbe V, Wincker P, Kanamori H, et al. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science, 345(6194):1251788–1251788.

Mora F, Castillo D, Lado B, Matus I, Poland J, Belzile F, von Zitzewitz J, and del Pozo A. 2015. Genome-wide association mapping of agronomic traits and carbon isotope discrimination in a worldwide germplasm collection of spring wheat using SNP markers. Molecular Breeding, 35: 69.

Moschen S, Bengoa Luoni S, Paniego NB, Hopp HE, Dosio GAA, Fernandez P, Heinz RA. 2014. Identification of Candidate Genes Associated with Leaf Senescence in Cultivated Sunflower (Helianthus annuus L.). PLoS One, 9(8): e104379.

Mueller ND, Gerber JS, Johnston M, Ray DK, Ramankutty N, Foley J. 2012. Closing yield gaps through nutrient and water management. Nature, 490(7419): 254–7.

Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, Gupta N, Neale BM, Daly MJ, Sklar P, Sullivan PF, Bergen S, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Purcell SM, Haas DW, Liang L, Sunyaev S, Patterson N, de Bakker PIW, Reich D, Price AL. 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nature Genetics, 44: 631–635.

Pei YF, Li J, Zhang L, Papasian CJ, Deng HW. 2008. Analyses and comparison of accuracy of different genotype imputation methods. PloS One, 3(10): e3551.

Poland JA, Brown PJ, Sorrells ME, Jannink JL. 2012a. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7(2): e32253.

Poland JA, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink JL. 2012b. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. The Plant Genome, 5(3): 103-113.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M a R, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81:559–575.

R Development Core Team. 2014. R: A language and environment for statistical computing, reference index version 3.1.3. Disponible en: http://www.R-project.org (verificado 05 junio 2015). R Foundation for Statistical Computing, Vienna, Austria.

Reynolds M, Foulkes J, Furbank R, Griffiths S, King J, Murchie E, Parry M, Slafer G. 2012. Achieving yield gains in wheat. Plant, Cell & Environment, 35, 1799–1823.

Röder MS, Huang XQ, Börner A. 2008. Fine mapping of the region on wheat chromosome 7D controlling grain weight. Functonal & Integrative Genomics 8(1): 79–86.

Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, McMullen MD, Holland JB, Buckler ES, Gardner CA. 2013. Comprehensive genotyping of the USA national maize inbred seed bank**.** Genome Biology, 14:R55.

Rutkoski JE, Poland J, Jannink J-L, Sorrells ME. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda) 3:427–39.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics, 78:629–644.

Sibson R. 1963. SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal, 30–34.

Singh RP, Hodson DP, Huerta-espino J, Jin Y, Njau P, Wanyera R, Herrera-Foessel SA, Ward RW. 2008. Will Stem Rust Destroy the World's Wheat Crop?. Advances in Agronomy, 98: 271-309.

Slafer GA, Araus JL. 2005. Physiological traits for imrpoving wheat yield under a wide range of conditions. Scale and Complexity in Plant Systems Research: Gene-Plant-Crop Relations, 21: 147–156.

Szűcs P, Blake VC, Bhat PR, Chao S, Close TJ, Cuesta-Marcos A, Muehlbauer GJ, Ramsay L, Waugh R, Hayes PM. 2009. An Integrated Resource for Barley Linkage Map and Malting Quality QTL Alignment. The Plan Genome, 2:134-140.

Tyagi S, Mir RR, Balyan HS, Gupta PK. 2014. Interval mapping and meta-QTL analysis of grain traits in common wheat (Triticum aestivum L.). Euphytica, 201(3): 367–380.

van Eeuwijk FA, Bink MCAM, Chenu K, Chapman SC. 2010. Detection and use of QTL for complex traits in multiple environments. Current Opinion of Plant Biology, 13: 193–205.

Windels CE. 2000. Economic and social impacts of fusarium head blight: changing farms and rural communities in the northern great plains. Phytopathology 90: 17–21.

Yu J, Buckler ES. 2006. Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology, 17: 155–60.

Yu J, Pressoir G, Briggs WH, Bi IV, 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: 203–208.

Zhu C, Gore M, Buckler ES, Yu J. 2008. Status and Prospects of Association Mapping in Plants. The Plant Genome, 1:5-20.