UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE AGRONOMÍA

EFICACIA RELATIVA DE MODELOS DE GWAS PARA VARIABLES ORDINALES

por

Agustín Mario GONZÁLEZ REYMÚNDEZ

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

Montevideo
URUGUAY
diciembre 2013

Tesis aprobada por el tribunal integrado por el Dr. Jorge Franco, la Dra. Mónica Balzarini y el Dr. Pablo Speranza, el 19 de diciembre de 2013. Autor: Lic. Agustín González, Directora Ing. Ag. (Ph.D) Lucia Gutierrez, Co-director Ing. Ag. (Ph.D) Ariel Castro

AGRADECIMIENTOS

Agradezco en primera instancia a mis directores de tesis, Ariel Castro y Lucía Gutiérrez, por su confianza, paciencia y tutoría en el entrenamiento y la concreción de todo este proceso. También agradezco a los compañeros y amigos del Departamento de Biometría, Estadística y Cómputo, por aconsejarme oportunamente y compartir su trabajo y experiencia. Por último, a las demás personas (familiares, amigos y otras víctimas circunstanciales), el mayor de mis reconocimientos.

TABLA DE CONTENIDO

Página

RESUMEN

El mapeo asociativo a escala genómica (GWAS) es uno de los métodos más eficaces para detectar genes de efecto cuantitativo (QTL, *quatitative trait loci*) que pueden ser utilizados para la selección asistida por marcadores en el desarrollo de cultivares. Los modelos convencionales para el estudio de QTLs son variaciones del modelo lineal mixto y por lo tanto suponen que los residuales asumen valores en una escala continua. No obstante, muchas de la variables fenotípicas de interés con base genética cuantitativa (eg. respuesta a enfermedades, supervivencia, conteos, fenología) son medidos en una escala ordinal. Por lo tanto, utilizar modelos lineales mixtos generales ignorando la naturaleza ordinal de las variables podría resultar en un compromiso de los resultados obtenidos a partir del GWAS.Si bien existen aproximaciones para tratar con variables medidas en una escala ordinal en la detección de QTLs (eg. transformaciones o modelos lineales generalizados), no está claro cuál es su ganancia relativa en la detección y estimación de efectos de QTLs. Para responder esto, mediante datos simulados y reales de genotipos de cebada se compararon cinco métodos de GWAS para variables ordinales en términos de su eficacia en la detección y estimación de efectos de QTLs, incluyendo modelos lineales generalizados y transformaciones de las variables originales. Para un rango muy amplio de tamaños de poblaciones, número de QTL y heredabilidades, no se encontraron diferencias entre métodos para la potencia y tasa de falsos descubrimientos en la detección de QTLs y las medidas del sesgo en las estimaciones de los efectos, fueron similares. Esto sugiere que la elección del método para tratar con variables ordinales no tiene un impacto mayor en la detección de QTLs a través de GWAS.

Palabras clave: Mapeo Asociativo, variables no-Gausianas, comparación de modelos, modelos lineales generalizados

RELATIVE EFFICACY OF GWAS MODELS FOR ORDINAL VARIABLES

SUMMARY

Genome wide association mapping (GWAS) is one of the most effective ways to detect quantitative trait loci (QTL) for marker assisted selection used to develop new cultivars. Conventional GWAS models for QTL detection are variation of the linear mixed model and therefore assume that residuals follow a normal distribution. However, many of the phenotypic variables of interest with a quantitative genetic basis (i.e. disease resistance, survival rate, counts, or phenology) are measured on an ordinal scale. Using general linear mixed models ignoring the ordinal nature of the variables could therefore compromise the results from GWAS. While there are approaches to deal with variables measured on an ordinal scale (i.e. transformations or generalized linear models), it is not clear what their relative gain in both the detection power and effect estimation is. To answer this, five methods for GWAS for ordinal variables, including generalized linear models and transformations were compared in terms of their relative efficacy in QTL detection and estimation. For a wide range of population sizes, number of QTLs, and heritabilities, no differences in power and false positives rate were detected across methods while similar bias were obtained for all methods. This suggests that the choice of the method for dealing with ordinal variables does not have a major impact on GWAS results.

Keywords: Association Mapping, non-Gaussian, model comparisons, generalized linear models

1. INTRODUCCIÓN

El mapeo asociativo a escala genómica (GWAS) permite la identificación de regiones genómicas que afectan la expresión de variables fenotípicas cuantitativas (*Quantitative Trait Loci*, QTL) a partir de poblaciones naturales o colecciones de bancos de germoplasma (Pritchard y Przerworski, 2001; Jannink y Walsh, 2002). Al igual que los métodos clásicos de mapeo de QTLs, el GWAS se basa en la detección de marcadores moleculares en desequilibrio de ligamiento (LD) con los QTLs (Pritchard y Przerworski, 2001; Jannink y Walsh, 2002). Esta detección sobre poblaciones diversas, permite explorar una mayor base genética, con mayor resolución y con menores tiempos y costos, que a partir de cruzamientos biparentales (Abdurakhmonov y Abdukarimov, 2008).

Si bien las fuentes de desequilibrio de ligamiento (LD) no debido a la distancia física entre marcadores (originado por estructura poblacional, selección natural y artificial, cuellos de botella, coancestría, etc.) (Jannink y Walsh, 2002; Pearson y Manolio, 2008) pueden provocar asociaciones entre marcadores y fenotipo, esto es corregido por los modelos convencionales de GWAS. Estos modelos remueven asociaciones espurias incluyendo covariables que representan el efecto de la estructura poblacional (Pritchard et al., 2000; Price et al., 2006), la coancestría (Yu et al., 2006), o una combinación de ambas (Yu et al., 2006; Malosetti et al., 2007; Kang et al., 2008).

No obstante, muchas variables fenotípicas de interés (eg. respuesta a enfermedades, supervivencia, conteos, fenología) son medidas en una escala ordinal, lo que podría implicar un compromiso en la validez de los resultados al ajustar estos modelos, que suponen un rango continuo para las variables fenotípicas (Madden y Hughes, 1995). El no-cumplimiento de supuestos distribucionales en las pruebas de asociación podría afectar la calidad de la inferencia sobre la posición y efecto de los QTLs, por lo que se han propuesto varias aproximaciones para tratarla, como son las transformaciones (Yang et al., 2006) y los modelos generalizados (Harville y Mee, 1984; Hackett y Weller, 1995; Spyrides-Cunha et al., 2000; Setakis et al., 2006).

Si bien se conocen las ventajas teóricas de aplicar estos métodos, no está claro cuál es la ganancia relativa en la detección de QTLs en el GWAS, al compararse empíricamente con métodos a priori más sencillos, en situaciones en las que no se conoce el modelo exacto que vincule la expresión fenotípica con los efectos genéticos. Para responder cuál es la ganancia de usar distintos métodos para el GWAS para variables ordinales, se compararon distintos modelos lineales generalizados (para modelar la no normalidad explícitamente) y transformaciones (para aproximar la distribución de la variable fenotípica a una normal). Para esto, el trabajo se dividió en dos secciones, que se presentan a continuación como dos artículos independientes.

En el primer artículo, a ser presentado ante la revista “Statistical Application in Genomic and Molecular Biology”, los métodos de GWAS para variables ordinales, se compararon sobre datos obtenidos *in silico* a partir de una matriz genotípica simulada y otra real. Esta sección se presenta según el formato exigido por la citada revista en idioma inglés.

En el segundo artículo, a ser presentado ante la revista “Agrociencia”, los métodos de GWAS se compararon a partir de datos fenotípicos simulados y reales sobre un panel de genotipos de Cebada. Esta sección se presenta con el formato exigido por la revista citada en idioma español.

Por último, luego de los capítulos correspondientes a ambos artículos, se presenta un capítulo dedicado a la discusión y conclusiones generales de todo el trabajo, seguido de la bibliografía general, consultada en ambos artículos.

2. PERFORMANCE OF SIMPLE GWAS MODELS FOR ORDINAL CHARACTERS

Agustín González-Reymúndez*[1], Ariel Castro[2], Lucía Gutiérrez[1]

[1]Departamento de Biometría, Estadística y Cómputo, Facultad de Agronomía, Garzón 780, Montevideo 12900, Uruguay

[2]Departamento de Producción Vegetal, Est. Exp. "Dr. Mario A. Cassinoni", Facultad de Agronomía, Universidad de la República. Ruta 3, Km. 373, Paysandú 60000, Uruguay

*Corresponding author (agugonrey@fagro.edu.uy)

2.1. ABSTRACT

Genome wide association mapping (GWAS) represents one of the most effective ways to detect quantitative trait loci (QTL). Conventional GWAS models assume that phenotypic variables fall on a continuous scale. When those models are applied to ordinal data, for example, GWAS results might be compromised by out of range predictions, heterogeneity of variances and inaccurate hypothesis testing results. There are several alternatives for mapping QTL of ordinal data through GWAS that include data transformation, generalized linear model, and non-parametric models). However, the empirical benefit of one method over another is not entirely clear. To answer this, five methods for GWAS for ordinal variables were compared in terms of their relative performance to detect QTL using simulated and real data. On a wide range of population sizes, number of QTL and heritabilities, no differences for power and false discovery rate were detected across methods, while similar bias in the estimation of QTL effects was obtained for all methods. This suggests that a classic GWAS model for ordinal data could still be appropriate for QTL mapping without compromising the efficacy of QTL detection.

Key words: Association Mapping, Non-Gaussian, model comparisons, generalized linear models

2.2. INTRODUCTION

Genome wide association mapping (GWAS) allows the identification of genomic regions controlling quantitative traits in diverse germplasm arrays (Quantitative Trait Loci, or QTL) (Pritchard and Przerworski, 2001; Jannink and Walsh, 2002). The use of diverse lines from natural populations or germplasm collections could produce higher accuracy and precision and less time and cost than biparental mapping populations and their results may be used directly to assist selection in the development of new cultivars (Abdurakhmonov and Abdukarimov, 2008).

Physical distance is not the only mechanism that generates linkage disequilibrium (LD). Evolutionary mechanism such as population admixture, selection, epistasis or coancestry may cause LD, increasing false positive rate in GWAS (Jannink and Walsh, 2002). Standard GWAS statistical models remove false LD caused by population structure and/or genetic relatedness like in Yu et al. (2006), Malosetti et al. (2007), Price et al. (2006), Pritchard, et al. (2000) , Zhao et al. (2007) and Kang et al. (2008). Being GWAS models variations of the linear mixed model, they assume normality of residuals (Henderson, 1984). When this assumption does not hold, inference on QTL position and effects could be negatively affected, causing a bias in the QTL estimated effect, out-of range predictions, or inaccurate hypothesis tests results (Casella and Berger, 1990; Wu et al., 2010). However, some of the relevant traits to map are not normally distributed (i.e. disease resistance, water deficit, and grain quality which are ordinal variables). To consider non-normality distributed variables, several approaches for QTL mapping have been implemented, such as generalized linear models, transformations, and bayesian and non-parametric models, (Spyrides-Cunha et al., 2000; Diao and Lin, 2006; Iwata et al., 2009; Coppieters et al., 1998). Additionally, it is not always intuitive which non-Gaussian model to use in every case because the genetic model causing the underlying distribution is not always evident. Although non-Gaussian models have been used for QTL mapping, there is no empirical evidence that these models perform better than Gaussian models. Most tests of significance are robust to deviations from normality and therefore, QTL detection should not be hindered by non-Gaussian traits. The

objective of this paper was to compare several GWAS models for studying ordinal variables in terms of their precision and accuracy in a wide range of simulated population scenarios. Power, false discovery rate, and bias in QTL effects were compared among methods to detect QTL using simulated and real genotypic data. Real genotypes were provided by the FONTAGRO project "Identification and Utilization of Durable Resistance to Barley Diseases in Latin America" (FONT 0617). Individuals consisted of 339 advanced barley breeding lines from programs of Latin America and ICARDA genotyped with 1560 SNPs markers from the Barley OPA (Close et al. 2009).

2.3. MATERIALS & METHODS

The general approach used in this paper was the comparison of five simple methods for GWAS of ordinal variables, including transformations and generalized linear models. Genotypic data consisted of two sets: one of *in silico* simulated data and one of real genotypic data. A phenotypic response (i.e an ordinal variable) was simulated for both sets. Performance of GWAS was compared across methods.

2.3.1. Populations and Genotypic data

Genotypes were simulated to represent an autogamous species like barley with two population sizes (n=100 and n=300). A number of 1200 binary markers were sampled every 1 cM to conform 70 windows of linked markers, where adjacent markers had a recombination frequency of 0.01 to ensure enough correlation between QTL and flanking markers. This was implemented in R (R Development Core Team, version 3.20, 2013) using code written specifically for this paper.

Real population data consisted of 339 genotypes of barley from breeding programs of Latin America and ICARDA/CIMMYT (Mexico) provided by the FONTAGRO project 'Identification and Utilization of Durable Resistance to Barley Diseases in Latin America' (FONT 0617) and cited elsewhere (Gutierrez et al., in prep). Briefly, 1560 SNPs were obtained following Illumina Golden Gate Bead Array Technology protocols (Illumina, San Diego, CA, USA). In order to reduce errors in the measurement of the markers, alleles that showed more than 10% missing data were removed. Additionally, all markers with a minor allele frequency of 10% were also excluded from the analysis, yielding 1096 usable SNPs. The estimated SNP position was based on the consensus map developed by Close et al. (2009). A cluster analysis within each chromosome was performed to identify an optimum number of independent groups of markers. This was implemented with the *fpc* package (Hennig, 2013) in R software (R Development Core Team, 2012).

For both sets of simulated and real genotypes, the phenotypic response were simulated by combining different values on QTL number (q = 3 and q = 9), heritability (h^2=0.5 and h^2 = 0.9), number of ordinal categories (k =2 and k = 10), and the assumed underlying genetic model (i.e. the function linking the phenotypic

response with the QTL states and effects; g = Normal and g = Multinomial). For each combination of these values, 16 alternative scenarios were obtained, each one as a (q, h^2, k, g) vector in which a model for the phenotypic response was specified. Additionally, for the simulated genotypic set, two n scenarios were evaluated.

2.3.2. Phenotypic data

For both simulated and real sets, phenotypic vector (**y**) was simulated to represent four distinct and relevant cases of ordinal responses. These cases were represented as a combination on g and k values. QTL effects position were taken at random from the set of markers, and the effects were sampled from a beta distribution with 1.6 and 5 as the shape parameters values. This was done to sample QTL with both big and very small effects. In the first two cases, to avoid a simplistic deterministic model and to include heritability considerations, each individual's phenotypic value was drawn from a normal distribution with a mean given by the linear combination of allelic states and effects, while the variance was determined as a function of the genetic variance and desired heritability ($\sigma^2_{error} = \sigma^2_{genetic}\ (1-h^2)/h^2$). By doing this, we created a continuous range of phenotypic distributions similar to the one assumed by the classic infinitesimal model. In the first case, a threshold was assumed to separate the range of **y*** in two categories (k=2) to define **y** as a binary ordinal response (i.e. presence/absence of disease). In the second case, with the same Gaussian underlying distribution, the range of **y*** was divided in 10 categories (k=10) to define **y** as an ordinal response (i.e. proportion of leaf area that is infected by a pathogen).

The third and fourth cases assumed a multinomial latent variable **y***, with the number of categories equal to n, the event probability equal to the normalized linear combination of QTL states and effects for each genotype, and the number of trials inversely proportional to h^2. By doing this, we created a discrete range similar to the one assumed by a Mendelian model. As for the first two cases, the range of **y*** was divided in two and ten categories, to compare all the cases among them. For each (q, h^2, k, g) vector, 1000 simulations of **y** values were done. En each simulation, the GWAS and model comparisons were performed

2.3.3. Statistical analysis

Five methods grouped in two models were defined to perform the GWAS for each simulation scenario. Population structure was determined by an Eigen-value decomposition of the standardized genotypic matrix and the selected axed were included in the model as fixed effects, as Price et al. (2006). To avoid over-parameterization, selection of Eigen-vectors was conducted by the elbow rule. The following general linear models were fitted to all simulation scenarios results:

$$f(\mathbf{y}) = \mathbf{Xb} + \mathbf{Sv} + \mathbf{e}$$

where f(y) correspond to a classic transformation of the response variable: identity, or no-transformation (**WT**), squared root transformation (**ST)**, and logarithmic transformation (**LT)**; **X** is the genotypic matrix, **b** is the vector of parametric marker effects, **S** correspond to a matrix with the selected eigen-vectors, **v** is the vector of population structure effects, and **e** is the vector of residual errors.

Additionally, generalized linear models were also fitted to all simulation scenarios results. Taking the values of **y** as ordered categories, both, a cumulative regression (k=10) and a binary regression (k=2) were performed, following Agresti (1996). The fitted model was:

$$g^{-1}(\mathbf{y}) = \mathbf{T\alpha} + \mathbf{Xb} + \mathbf{Sv}$$

where g^{-1} is a function that links the expected value of the response to a hypothetical distribution of the residuals or an inverse of the cumulative density function of the residual of a latent continuous variable, like in Agresti (1996). We defined two alternatives to the model, taking g as the normal distribution to perform cumulative (binary) probit regression (**CP**) and taking g as the logistic distribution, to perform cumulative (binary) logit regression (**CL**); **T** is an incidence matrix that allows to match each genotype with its own category given by **α**, that is the vector of parametric thresholds of the latent variable range.

For each of the five methods (**WT**, **ST**, **LT**, **CP** and **CL**) a marker by marker regression was conducted to obtain estimates of marker effects. Wald tests were performed to determine significance using a Bonferroni (1935) corrected p-value by adjusting with the number of independent tests. GWAS performance was analyzed as power, false discovery rate (FDR) and bias in QTL effect estimation, as in Arbelbide et al. (2006). Power and FDR were calculated as TP/(TP + FN) and as FP/(FP + TP) respectively. True positives (TP) were obtained as the number of marker groups including a true QTL with at least one significant marker-trait association. False negatives (FN) are the number of marker groups with a true QTL but with no significant marker-trait association. Finally, false positives (FP) are the number of the windows without a true QTL but with at least one significant marker-trait association. Bias in QTL effect estimation was calculated as the mean deviation of the true effect and the estimated effect, divided by the true effect. Each estimate was inverse transformed, according to the model used to calculate it to recover original units.

GWAS method comparison was visualized by comparing performance metrics distributions with boxplots and comparing medians with a Friedman test, taking methods as treatments and simulation scenarios (combinations of *q*, *h2*, *k* and underlying genetic model) as blocks. When a positive result was obtained by Friedman test, multiple-comparisons with a Nemenyi-Damico-Wolfe-Dunn test were conducted. Simulation, calculation of precision-accuracy metrics and methods comparisons were conducted in R (R Development Core Team, version 3.20, 2013) using functions from packages *stats*, *ordinal* (Christensen, 2012) and code written specifically for this paper.

2.4. RESULTS

2.4.1. Simulated genotypes

The ranges of empirical distributions for power and FDR overlap for all GWAS methods, for both k =2 (Fig. 1) and k = 10 (Fig. 2). All efficacy metrics were higher for large population sizes and heritabilities. Power, FDR and estimation bias increased at decreasing number of QTL (q). Heritability did not affect GWAS methods. In general, no interaction was detected between GWAS methods and population parameters. The Friedman test failed to detect significant differences for power and FDR across GWAS methods, having a median of 0.59 for power and 0 for FDR through simulated scenarios as vectors (n, q, k, h^2, g) (Table 1).

Table 1: Medians of power, FDR and bias of QTL effects for different GWAS methods as treatments and simulated populations scenarios as blocks: WT (simple linear regression without transformation), ST (simple linear regression on square root transformed data), LT (simple linear regression on logarithm transformed data), CP (cumulative simple regression with probit link) and CL (cumulative simple regression with logit link) for the simulated genotypic set (S.Geno) and the real genotypic set (R.Geno)

	Power		**FDR**		**Bias**[†]	
Method	**S.Geno.**	**R.Geno.**	**S.Geno.**	**R.Geno.**	**S.Geno.**	**R.Geno.**
LT	0.59	0.44	0	0.75	4.08[a]	1.04 [a]
CL	0.58	0.44	0	0.75	-0.97[b]	-0.04 [b]
CP	0.58	0.44	0	0.75	-0.00[bc]	-0.11b [c]
WT	0.59	0.44	0	0.75	-0.32[c]	-0.14 [c]
ST	0.59	0.44	0	0.75	-0.78[c]	-0.24 [c]

† Friedman test detected significance differences at p=0.05. Different letters indicates differences between GWAS methods performed with Nemenyi-Damico-Wolfe-Dunn test

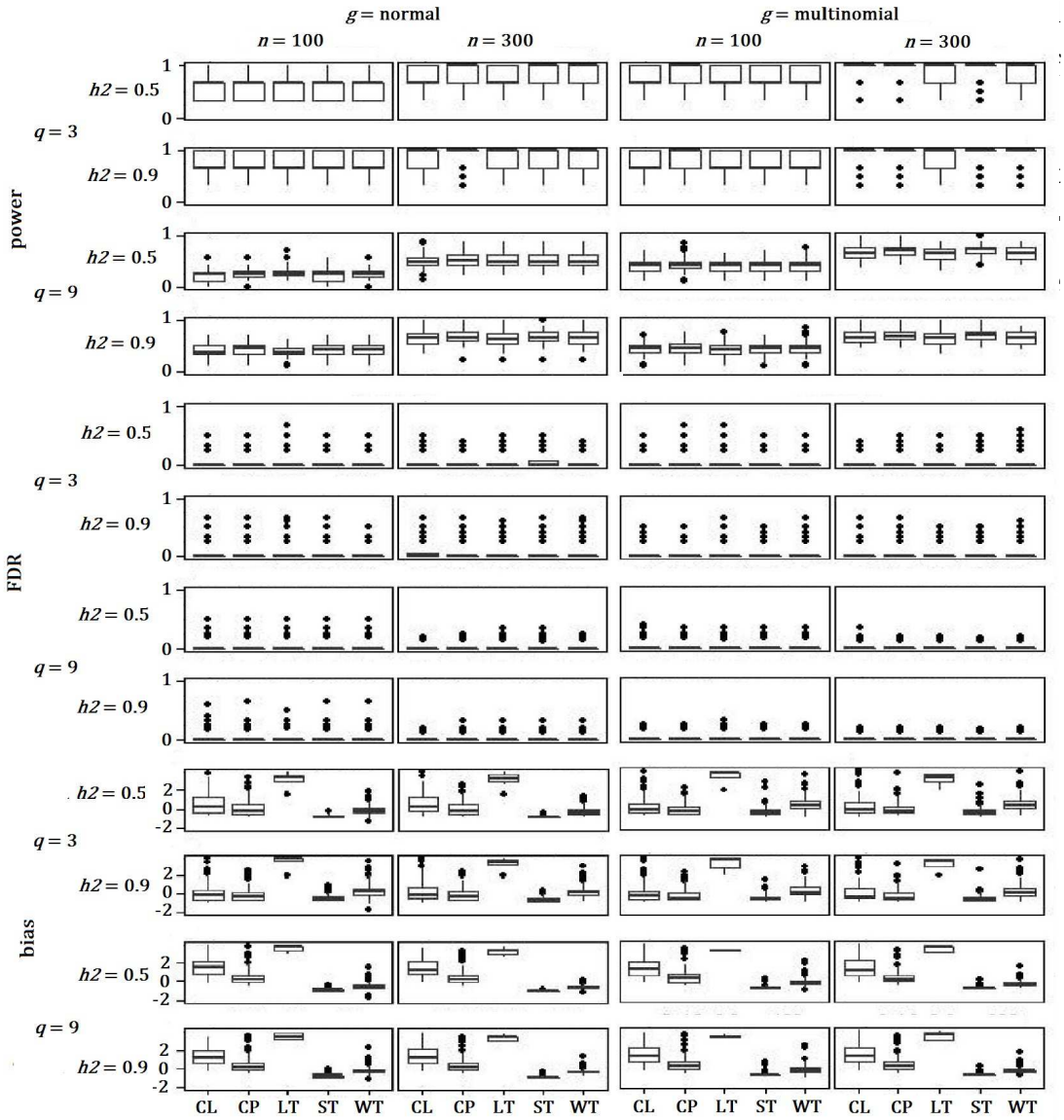

Figure 1: Power and FDR of QTL detection and bias of QTL effect for simulated genotypic set with ten categories of an ordinal response and different number of QTL (q), heritability (h^2), underlying genetic model (g) and population size (n), evaluated with five different methods: WT (simple linear regression without transformation), ST (simple linear regression on square root transformed data), LT (simple linear regression on logarithm transformed data), CP (cumulative simple regression with probit link) and CL (cumulative simple regression with logit link)

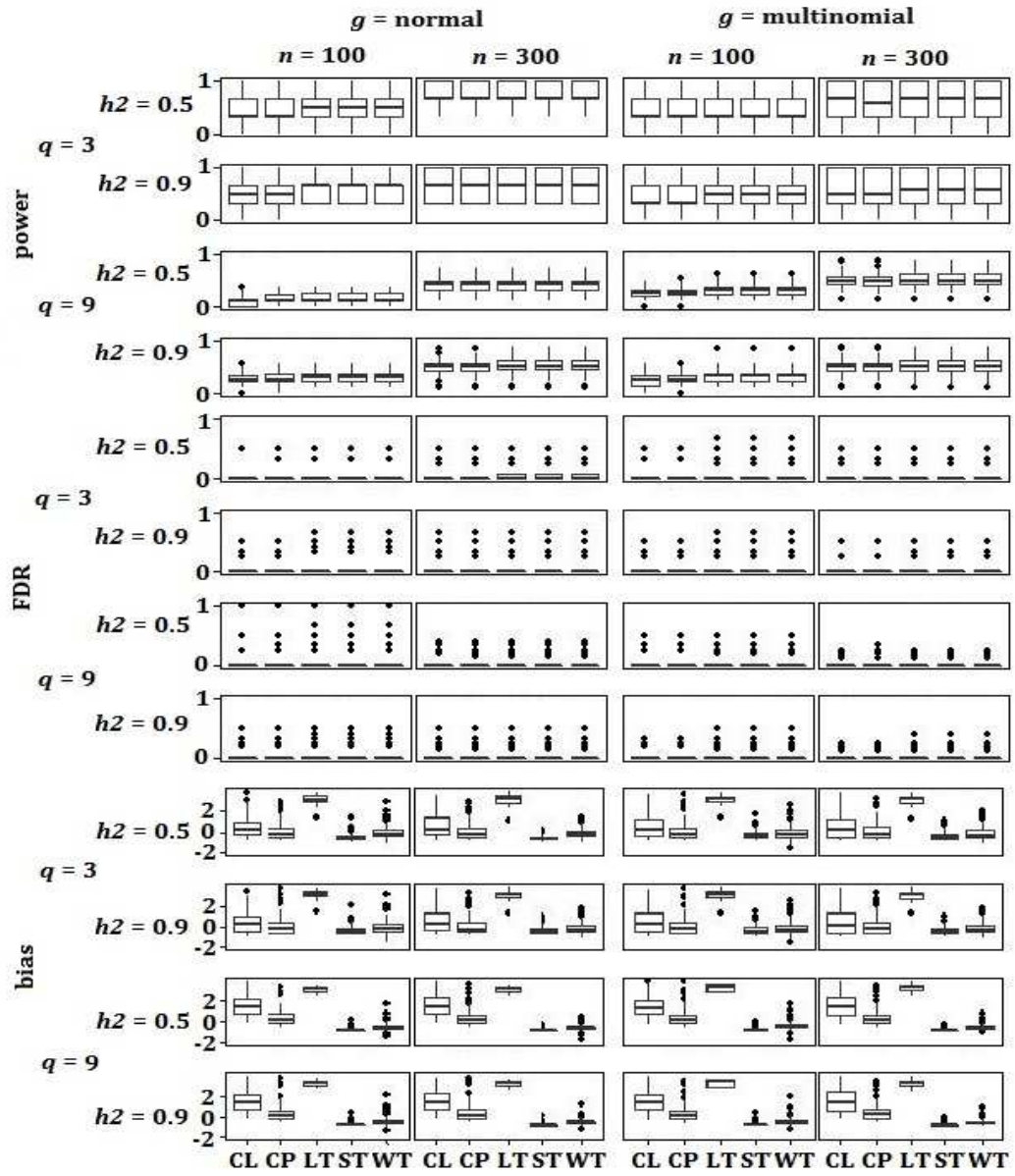

Figure 2: Power and FDR of QTL detection and bias of QTL effect for simulated genotypic set with two categories of an ordinal response and different number of QTL (q), heritability (h^2), underlying genetic model (g) and population size (n), evaluated with five different methods: WT (simple linear regression without transformation), ST (simple linear regression on square root transformed data), LT (simple linear regression on logarithm transformed data), CP (cumulative simple regression with probit link) and CL (cumulative simple regression with logit link

2.4.2. Real genotypes

We used the first two principal component axes to correct for population structure because they showed an evident clustering of genotypes in 3 groups (Fig. 3). We did not detect any significant differences in power or FDR across GWAS methods. Because of the apparent interaction between GWAS methods for power (Fig. 4), a Friedman test for each level of *g* and *k* was performed but no significant difference was detected (data not shown). For FDR, no significant differences were detected with a median value of 0.75 trough GWAS methods.

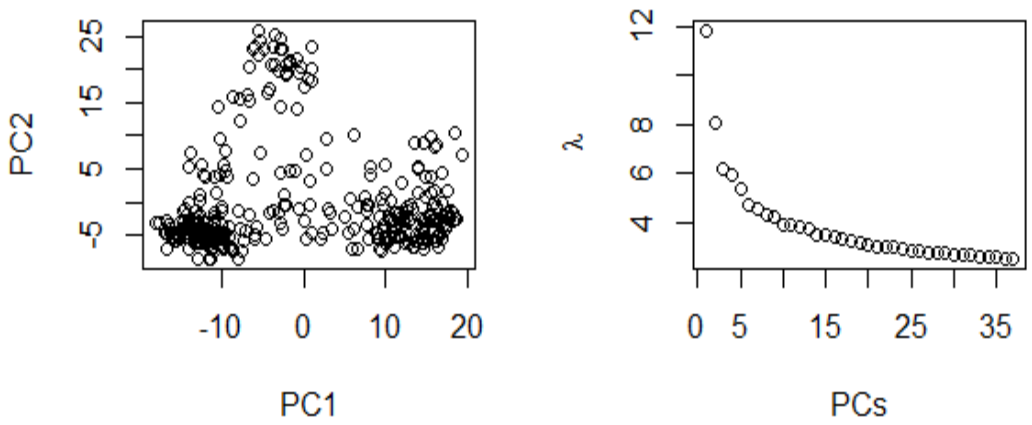

Figure 3: Principal component analysis of genotypic data: plot of the two first components (left) and plot of eigenvalues by each component (right). PC1: principal component 1, PC2: principal component 2, PCs: principal components, λ: eigenvalues

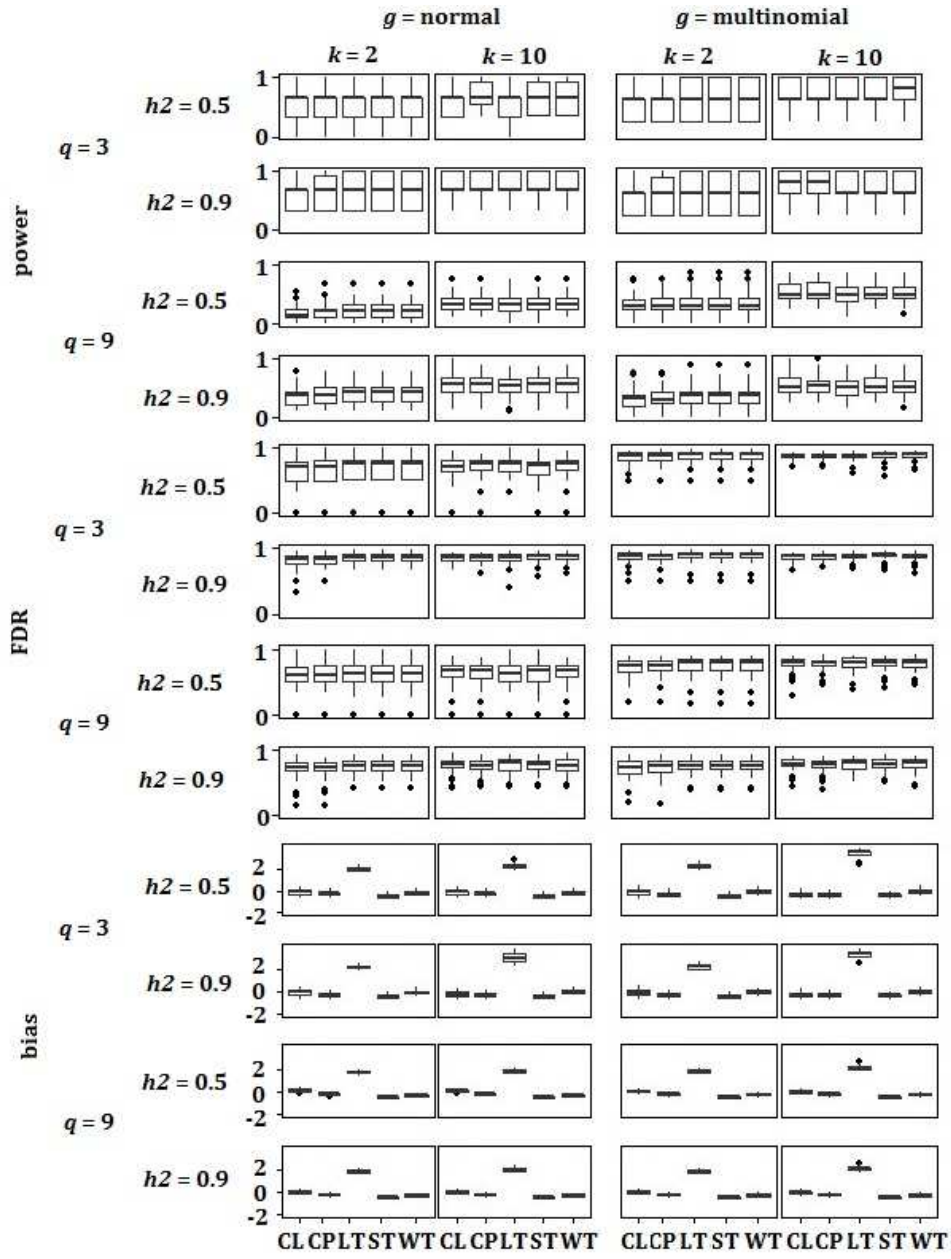

Figure 4: Power and FDR of QTL detection and bias of QTL effect for real genotypic set with different number of QTL (q), heritability (h2), underlying genetic model (g) and number of ordinal categories (k), evaluated with five different methods: WT (simple linear regression without transformation), ST (simple linear regression on square root transformed data), LT (simple linear regression on logarithm transformed data), CP (cumulative simple regression with probit link) and CL (cumulative simple regression with logit link).

The most biased method was LT. Because the apparent interaction between GWAS method and k (ie. LT method was higher for k=10 than for k=2), Friedman test was performed for each value of k, separately. For both values of k, WT, ST, CP and CL were no significant different between them and different respecting LT (Table 1).

2.5. DISCUSSION

In this work two approximations were used to test differences in empirical performance of GWAS for ordinal characters: transformations i.e. using the expected value of a function of **y** to try to approximate the distribution of **y** to normality, and generalized linear models, which uses a function of the expected value of **y** and link it to a hypothetical distribution of **y**. Both types of approximation provided a simple way to compare the relative performance of the GWAS methods.

We found no evidence of differences across GWAS methods for both power and FDR ($p<0.05$). The most unbiased method was the **CL** for simulated genotypes; however, for real genotypes no significant differences occurred between **CL** and **WT**, **ST,** and **CP** methods. This suggest that models assuming normality performed equally well as models that explicitly took into account the ordinal nature of the variable **y**.

For simulated data, FDR was zero in most cases. For real data, FDR was much larger, taking a median value of 0.75 across (q, h^2, k, g) vectors. This implies a strong effect of the population structure despite the very stringent threshold used for correcting p-values (that explain the pattern in simulated data). A more efficient mixed model to correct for population structure could decrease these values. However, the purpose of this work was the comparison of different GWAS methods for ordinal data. The GWAS methods compared had the same power and FDR in QTLs detection and similar values for estimation bias. This was obtained for a wider range of heritabilities (including a $h^2 = 0.1$, data not shown). However, the very low error comparison rate we used could have masked the potential differences under some population scenarios.

Eventually, these results can be contrasted with previous work on GWAS for ordinal variables where QTLs were reported with models which did not consider discretized phenotypes. Further work should be done to compare a wider spectrum of methods at varying error comparison rates, lower heritabilities and correction for population structure with efficient mixed models.

2.6. ACKNOWLEDGEMENTS

We would like to thank Dr. Pablo Speranza, Dr. Mónica Balzarini, Dr. Silvia Germán, Dr. Jorge Franco and Dr. Marcos Malosetti, who made valuable suggestions and corrections on initial drafts. Finally, we would to express our gratitude to the the FONTAGRO project “Identificación y utilización de resistencia durable a enfermedades de cebada en américa latina” members who provide us with the real data that were used in this work**.** Funding for this research has been provided by Fondo Regional de Tecnología Agropecuaria (FONTAGRO) Grant 0617-06, Universidad de la República, Uruguay (CSIC Grupos) and Agencia Nacional de Investigacion e Innovacion (ANII_3546) graduate scholarship awarded to A. Gonzalez.

2.7. REFERENCES

1. Abdurakhmonov, I. Y. and A. Abdukarimov (2008): "Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources", Int. Journ. of P. Genom. 2008, 1-18.
2. Agresti, A. (1996): An Introduction to Categorical Data Analysis, A. Agresti-John Wiley & Sons.
3. Arbelbide, M., J. Yu and R. Bernardo (2006): "Power of mixed-model QTL mapping from phenotypic, pedigree and marker data in self-pollinated crops", Theor. App. Genet., 112, 876–84.
4. Bonferroni, C.E. (1935): “Il calcolo delle assicurazioni su gruppi di teste”. *In*: Studi in Onore del Professore Salvatore Ortu Carboni.
5. Casella, G. and R. Berger (1990). Statistical inference, Duxbury.
6. Close T. J., P. R. Bhat, S. Lonardi, Y. Wu, N. Rostoks, L. Ramsay, A. Druka, N. Stein, J. T. Svensson, S. Wanamaker, S. Bozdag, M. L. Roose, M. J. Moscou, S. Chao, R. K. Varshney, P. Szücs, K. Sato, P. M. Hayes, D. E. Matthews, A. Kleinhofs, G. J. Muehlbauer, J. DeYoung, D. F. Marshall, K. Madishetty, R. D. Fenton, P. Condamine, A. Graner and R. Waugh (2009): “Development and implementation of high-throughput SNP genotyping in barley”, BMC Genom., 10, 1-13.
7. Christensen, R. H. B. (2012): “Ordinal Regression Models for Ordinal Data”. R package version 2012. 01-19.
8. Coppieters, W., A. Kvasz, F. Farnir, J. J. Arranz, B. Grisart, M. Mackinnon and M. Georges (1998): "A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sib pedigrees: application to milk production in a granddaughter design", Genetics, 149, 1547–55
9. Diao, G. and D.Y.Lin. (2006): "Improving the power of association tests for quantitative traits in family studies", Genet. Epidemiol., 30, 301–313.
10. Friedman, M. (1937): “The use of ranks to avoid the assumption of normality implicit in the analysis of variance”, J. Am. Statist. Assoc., 32: 675-701.
11. Henderson, C. R. (1984). Applications of linear models in animal breeding. University of Guelph.

12. Hennig, C. (2013): “Fpc: Flexible procedures for clustering”. R package version 2.1-5.
13. Iwata, H., K. Ebana and S. Fukuoka. (2009): “Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among *Oryza sativa L.* germplasms”. Theor. Appl. Genet., 18: 865-880.
14. Jannink, J. and B. Walsh. (2002): "Association mapping in plant populations", Quantitative Genetics, Genomics and Plant Breeding, M. S. Kang-CABI.
15. Kang, H. M., N. Zaitlen, C. Wade, A. Kirby, D. Heckerman, M.J. Daly and E. Eskin (2008): "Efficient control of population structure in model organism association mapping", Genetics, 178, 1709–1723.
16. Madden, L.V. and G. Hughes (1995): “Plant disease incidence: distributions, heterogenity, and temporal analysis”, Ann. Rev. Phytopathol, 33, 529-564.
17. Malosetti, M., C. van der Linden, B. Vosman and F. van Eeuwijk (2007): "A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato", Genetics, 175, 879–889.
18. Nemenyi, P. B. (1963): “Distribution-free Multiple Comparisons”, Princeton: Princeton University.
19. Price, A. L., N. Patterson and R. Plenge (2006): "Principal components analysis corrects for stratification in genome-wide association studies", Nature , 38, 904–909.
20. Pritchard J. K. and M. Przeworsk (2001): “Linkage disequilibrium in humans: models and data”, Am. J. Hum. Gen., 69:1-14.
21. Pritchard, J. K., M. Stephens P. Donnelly (2000): "Inference of population structure using multilocus genotype data", Genetics, 155, 945–959.
22. R Development Core Team (2012): “R: A language and environment for statistical computing”. R Foundation for Statistical Computing.

23. Spyrides-Cunha, M., C. Demétrio and L. Camargo. (2000): "Proportional odds model applied to mapping of disease resistance genes in plants", Genet. Mol. Biol., 23, 223-227
24. Wu, R., C. Ma and G. Casella (2010): Statistical genetics of quantitative traits: Linkage, maps and QTL. Springer.
25. Yu J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. Mcmullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovic and E. S. Buckler (2006): “A unified mixed-model method for association mapping that accounts for multiple levels of relatedness”, Nat. Genet., 38, 203-208.
26. Zhao, K., M. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, and P. Marjoram (2007): "An Arabidopsis example of association mapping in structured samples", PLoS Genet., 71-82.

3. COMPARACIÓN DE MODELOS DE ANÁLISIS PARA LA RESPUESTA A ENFERMEDADES EN CEBADA MEDIDA EN ESCALA ORDINAL

Agustín González-Reymúndez*[1], Ariel Castro[2], Lucía Gutiérrez[1]

[1]Departamento de Biometría, Estadística y Cómputo, Facultad de Agronomía, Garzón 780, Montevideo 12900, Uruguay. agugonrey@fagro.edu.uy

[2]Departamento de Producción Vegetal, Est. Exp. "Dr. Mario A. Cassinoni", Facultad de Agronomía, Universidad de la República. Ruta 3, Km. 373, Paysandú 60000, Uruguay

3.1. SUMMARY

GWAS Model Comparison in Barley for Disease Response on an Ordinal Scale

Diseases are one of the factors that cause larger economic losses in barley production. Like other attributes of agro-economic importance (eg. yield, grain protein or malt quality), the response to diseases in barley can be mapped with the same techniques used to detect quantitative effect genes (QTL). One of these techniques is the Genome Wide Association Studies (GWAS) that provide an efficient way to QTL detection. However, since disease response is typically measured on a discrete scale, fitting classical normal error models could imply a compromise in the validity of GWAS results. Though several alternatives to deal with the ordinal variables in the detection of QTLs (eg transformations, generalized linear models) exist, it is unclear how much gain is obtained in QTL detection and effect estimation by using these models. To answer this, five methods of GWAS for ordinal variables were compared in terms of their effectiveness in the detection of QTL and effects estimation, including generalized linear models and transformations. We used real and simulated phenotypic data for disease. No differences for power and false discovery rate were detected across methods, while similar bias in the estimation of QTL effects was obtained for all methods. This suggests that a classic GWAS normal error model could still be appropriate for QTL mapping of barley disease response.

Key words: GWAS, ordinal variables, models comparisons

3.2. RESUMEN

Las enfermedades son uno de los factores que causan mayores pérdidas económicas en el cultivo de cebada. Al igual que en otros atributos de importancia agro-económica (eg. rendimiento, proteína en grano o calidad maltera) la respuesta a enfermedades en la cebada puede mapearse con las mismas técnicas usadas para detectar genes de efecto cuantitativo. Una de estas técnicas es el Estudio de Asociación a través del Genoma (*Genome Wide Association Study*, GWAS). No obstante, dado que este tipo de respuesta suele medirse en una escala discreta, el ajuste de los modelos clásicos de errores normales implica un compromiso en la validez de los resultados del GWAS. Si bien existen varias alternativas para tratar con variables ordinales en la detección de QTLs (eg. transformaciones, modelos lineales generalizados), no está claro cuánta es la ganancia en la detección y estimación de efectos de QTLs. Para responder esto, cinco métodos de GWAS para variables ordinales se compararon en términos de su eficacia en la detección y estimación de efectos de QTLs, incluyendo modelos lineales generalizados y transformaciones, usando datos fenotípicos simulados y reales para la respuesta a enfermedades. No se encontraron diferencias entre métodos para la potencia y tasa de falsos descubrimientos en la detección de QTLs, y las medidas del sesgo en las estimaciones, fueron similares. Esto sugiere que un modelo clásico de GWAS de errores normales puedes seguir siendo adecuado para mapear QTL para la respuesta a enfermedades en cebada

Palabras clave: GWAS, variables ordinales, comparación de modelos

3.3. INTRODUCCIÓN

Las enfermedades son uno de los factores que causan mayores pérdidas en el cultivo de cebada, tanto en producción como en calidad, y su control es una tarea fundamental (HGCA, 2009). Al igual que en otros atributos de importancia agro-económica, como rendimiento, proteína en grano o calidad maltera, la respuesta a enfermedades en la cebada puede mapearse con las mismas técnicas usadas para detectar genes de efecto cuantitativo (Hayes et al., 2003). La localización y estimación de estos factores de resistencia permite la selección de variedades más resistentes a través de cruzamientos dirigidos, sin necesidad de fenotipado en cada ciclo de selección (Steffenson et al., 1996; Bertrand et al., 2008). Esto es ventajoso, dado que no siempre se dan las condiciones para el desarrollo de la enfermedad y por lo costoso del fenotipado (Steffenson et al., 1996; Bankina & Gaile, 2009).

Tradicionalmente, el mapeo de los QTLs vinculados a la respuesta a enfermedades se ha llevado a cabo mediante cruzamientos biparentales entre líneas puras resistentes y susceptibles (Hayes et al. 2003). Otra maneras de mapear QTLs, que presenta varias ventajas prácticas, es el mapeo asociativo a través del genoma (GWAS), que utiliza una muestra del germoplasma de referencia (Jannink y Walsh, 2002). Esto permite explotar una variabilidad genética mucho mayor y acceder a muchas más recombinaciones que a partir de un esquema de cruzamientos biparentales (Abdurakhmonov y Abdukarimov 2008). Al tener un número mucho mayor de recombinaciones, el decaimiento del desequilibrio de ligamiento (LD) es más pronunciado y se podría tener una mayor precisión en la identificación de los QTL (Jannink y Walsh, 2002; Roy et al., 2010). Como el GWAS se basa en la inferencia de asociaciones entre fenotipo y marcador, aquellas asociaciones que surjan por el LD no debido a distancia genética deben ser removidas de los análisis (Jannink y Walsh, 2002).

Los modelos convencionales de GWAS corrigen por los efectos del LD espurio de distinta forma y pueden clasificarse en modelos fijos, que corrigen por estructura poblacional a través de factores fijos (Pritchard et al., 2000; Price et al.,

2006); modelos *kinship*, que usan un modelo mixto animal modificado (Yu et al., 2006) y modelos mixtos, que corrigen por estructura y coancestría simultáneamente (Yu et al. 2006; Zhao et al. 2007; Malosetti et al., 2007; Kang et al., 2008). Sin embargo, en el caso de la respuesta a enfermedades, el fenotipado se evalúa mediante juicio visual del fitopatólogo, en una escala ordinal, como porcentajes en una escala discreta o en función de criterios cualitativos (Horsfall y Barratt, 1945; Xu y Atchley, 1996). Esto produce desvíos de la normalidad que podrían afectar la inferencia a partir de los modelos convencionales de GWAS (eg. predicciones fuera de rango, heterogeneidad de varianzas, asimetría) (Madden y Hughes 1995).

Se han utilizado distintas alternativas para realizar el GWAS para respuesta a enfermedades en cebada considerando errores no Gaussianos, en particular cuando el fenotipado se realiza en plántula en una escala binaria de resistente/no resistente (Castro et al., 2002; Roy et al., 2010; Wang et al., 2012). No obstante, no existen comparaciones empíricas de qué método es más eficaz para qué tipo de variable. Para responder esto, en este trabajo se compararon cinco métodos para el GWAS de variables ordinales, incluyendo modelos lineales generalizados y transformaciones. Esto fue llevado a cabo mediante el análisis de datos genotípicos de cebada de programas de América Latina y del ICARDA/CIMMYT (México), y datos fenotípicos de respuesta de los genotipos a las infecciones de mancha borrosa (causada por *Cochliobolus sativus*) y roya de la hoja (causada por *Puccinia hordei*), y simulaciones, a partir de la matriz de datos genéticos. Los métodos se compararon con el fin de determinar cuál es la ganancia al usar un modelo u otro según el tipo de variable fenotípica. Conocer qué método es mejor para la ubicación de factores de resistencia, así como la estimación de sus efectos podría tener un impacto inmediato en el desarrollo de variedades resistentes en el mejoramiento genético, como alternativa más económica y menos nociva para el ambiente y la salud humana que el uso de plaguicidas (Phipps y Park, 2002).

3.4. MATERIALES Y MÉTODOS

3.4.1. Material vegetal y fenotipado

Los datos genotípicos y fenotípicos reales fueron provistos por el proyecto "Identificación y utilización de resistencia durable a enfermedades de cebada en América Latina", financiado por el Fondo Regional de Tecnología Agropecuaria (FONTAGRO) (http://www.fontagro.org/proyectos/identificación-y-utilización-de-resistencia-durable-enfermedades-de-cebada-en-américa-lati). Se utilizaron un total de 360 genotipos de cultivares de programas de mejoramiento de América Latina y de ICARDA/CIMMYT (México). Los datos fenotípicos correspondieron a la evaluación de la severidad a la infección por mancha borrosa (*C. sativus*) y roya de la hoja (*P. hordei*) medidas en los estados de plántula y planta adulta para cuatro ambientes, como combinaciones de localidad (La Estanzuela, 57°41'W, 34°20'S, y la estación experimental Dr. Mario Cassinoni (EMAC) en Paysandú, 58°03'W, 34°20'S) y año (2009-2010). Para planta adulta, la severidad fue medida como porcentaje de área infectada y para plántula, en una escala de 0 a 4 descrita por Stakman et al. (1962).

3.4.2. Análisis de datos fenotípicos

Los datos fenotípicos se analizaron a través de los BLUP obtenidos al ajustar un modelo de Federer (Federer, 1961) y la notación siguió la de los trabajos Eckerman et al. (2001) y Verbyla et al. (2003):

$$Y_{ijk} = \mu + \beta_i + G_j + \varepsilon_{ijk},$$

donde y_{ijk} es la variable de respuesta (i.e severidad), μ es la media general, β_i es el efecto aleatorio del bloque incompleto i $\beta_i \sim N(0, \sigma_b^2)$, Gj es el efecto genotípico y $\varepsilon_{ij}k$ es el error. El modelo para Gj fue el siguiente:

$$G_j = g_j + c_{(i)k}$$

g_i es el efecto del i-ésimo genotipo con i=1,…,n_g (número de genotipos a probar); y $c_{(i)k}$ representó un efecto fijo para el k-ésimo testigo con j=n_g+1,…,n_g+n_c, con n_c igual al número de testigos. En los casos donde una verdadera repetición existió, un término para el bloque fue incluido. Además, utilizando los testigos repetidos una corrección espacial por fila y columna fue usada con diferentes estructuras de varianza-covarianza. El análisis se hizo mediante el procedimiento PROC MIXED de SAS Statistical Software (SAS Institute, 2004) para obtenera las mejores estimaciones insesgadas (BLUE). Todos estos análisis fueron llevados a cabo siguiendo lo descripto en Gutierrez et. al. (en prep.).

3.4.3. Datos genotípicos

Los datos genotípicos se arreglaron en una matriz de tamaño 360 (líneas) por 1560 (SNP). Con el fin de reducir errores en las valoraciones de alelos se excluyeron del análisis los marcadores que presentaron más de un 10% de datos faltantes y los marcadores con una frecuencia de alelo menor al 10% también fueron excluidos del análisis, reduciendo la matriz a dimensiones 336x1096. La posición estimada de los SNP se basó en el mapa consenso desarrollado por Close et al. (2009).

3.4.4. Datos simulados

El proceso de simulación de datos fenotípicos sobre la matriz genotípica real siguió el procedimiento de (González-Reymúndez 2013, en prensa), obteniendo los datos simulados muestreando marcadores al azar imponiendo sobre cada uno un efecto. Una vez definidos los QTL sobre la matriz genotípica real, se simularon valores fenotípicos haciendo variar el número de QTL (q=10 y q=14) (aproximadamente, los números de QTL encontrados para ambas enfermedades), heredabilidad (tres valores representando heredabilidades baja, media y alta: $h2$=0,1, $h2$=0,5, $h2$=0,9) y número de categorías (k=6 y k=20, el rango para la cantidad de categorías fenotípicas correspondientes a los datos reales utilizados). Los datos

fenotípicos se generaron suponiendo una variable latente **y*** con distribuciones, normal y multinomial, para cada valor de *h2 y q* (González-Reymúndez 2013, en prensa).

3.4.5. Análisis Estadístico

Siguiendo a González-Reymúndez 2013, para cada una de las variables fenotípicas (reales y simuladas) se ajustaron métodos de regresión simple sin transformación (WT), con transformación raíz cuadrada (ST), con transformación logaritmo (LT), regresión ordinal probit (CP) y regresión ordinal logit (CL). En todos los casos, los modelos se ajustaron para cada marcador y se efectuaron pruebas de Wald sobre los efectos estimados de cada marcador usando una corrección de Bonferroni (Bonferroni, 1935) usando el número de grupos de marcadores ligados, tomando 10 cM como distancia máxima entre pares de marcadores. Para controlar por el desequilibrio de ligamiento debido a la estructura poblacional, se utilizaron los scores correspondientes al análisis de componentes principales a partir de la matriz genotípica como efectos fijos y se usó la regla del codo para la selección de ejes.

Para las variables fenotípicas reales, los métodos de GWAS se compararon como en Yu et al. (2006) mediante q-q plots, graficando los p-valores de las pruebas de asociación correspondientes a marcadores no ligados a QTLs contra una distribución uniforme.

Para las variables fenotípicas simuladas, debido a que tanto los efectos como las posiciones de los QTLs fueron impuestas, los modelos se compararon en base a la potencia (P=TP/(TP+FN)), tasa de falsos descubrimientos (FDR= FP/(FP + TP) y al sesgo en la estimación de efectos de QTLs, donde TP son verdaderos positivos, FN son falsos negativos y FP son falsos positivos. Para calcular cada uno de ellos, se definieron ventanas de marcadores cuyas distancias genéticas entre sí no superara los 10 cM. Sobre el conjunto de estas ventanas, los TP se definieron como número de ventanas con un QTL real, en las cuales algún marcador resulta significativo para la prueba de asociación; los falsos negativos (FN) se definieron como el número de

estas ventanas dentro de las cuales ningún marcador resulta positivo y los falsos positivos (FP) se definieron como número de ventanas dentro de las cuales no hay QTL pero algunos marcadores resultan significativos. El sesgo se calculó como los desvíos promedio de las estimaciones con respecto a los valores impuestos para los efectos. Para llevar las estimaciones hechas por los distintos modelos a las mismas unidades, se realizaron las transformaciones inversas correspondiente a cada método sobre las estimaciones. La comparación de métodos se realizaron mediante las medianas de cada métrica luego de tomar 1000 iteraciones para cada combinación de valores de *q*, *h2*, *fy** y *k*. Sobre estos valores, se realizaron pruebas de Friedman (Friedman, 1937) para cada métrica, cómo un análogo no parametrico al análisis de varianzas para el diseño de bloques al azar. Cuando se obtuvieron diferencias significativas entre modelos, se realizó una prueba de comparaciones múltiples de Nemenyi-Damico-Wolfe-Dunn (Nemenyi, 1963).

3.5. RESULTADOS

3.5.1. Datos reales

Los componentes principales seleccionados para corregir por la estructura poblacional fueron los primeros dos (Fig. 3[1]). Los métodos de GWAS tuvieron un comportamiento en general equivalente para la respuesta a ambos tipos de infección (Fig. 5). Para roya de la hoja en plántula, los métodos se separaron en dos tendencias, una correspondiente a los modelos generales y la otra a los generalizados. En el caso de mancha borrosa en planta adulta, el método CL fue el más preciso.

En el cromosoma 4H se detectó un QTL entre 59 y 78 cM para roya de la hoja en planta adulta, para los modelos ST y LT. En el cromosoma 5H se detectaron 5 QTL: entre 15 y 35 cM para roya de la hoja en plántula (modelos WT y ST), entre 74,5 y 94,5 cM para roya de la hoja en planta adulta (WT, ST, CP, CL); entre 112 y 132 cM y entre 139 y 159,6 cM para mancha borrosa en planta adulta (todos los métodos). En el cromosoma 6H se detectaron 2 QTL: entre 38,70 y 58,75 cM para roya de la hoja en planta adulta (modelos ST, LT, CP) y entre 119 y 129 cM para roya de la hoja en plántula (modelos CP y CL). En el cromosoma 7 se detectaron 2 QTL: entre 0 y 10 cM para roya de la hoja en planta adulta, con todos los modelos excepto CL; entre 98,50 y 112,40 cM, para mancha borrosa en plántula (modelo CL).

1 Mismo resultado que en el capítulo anterior para el ACP de la matriz genotípica real

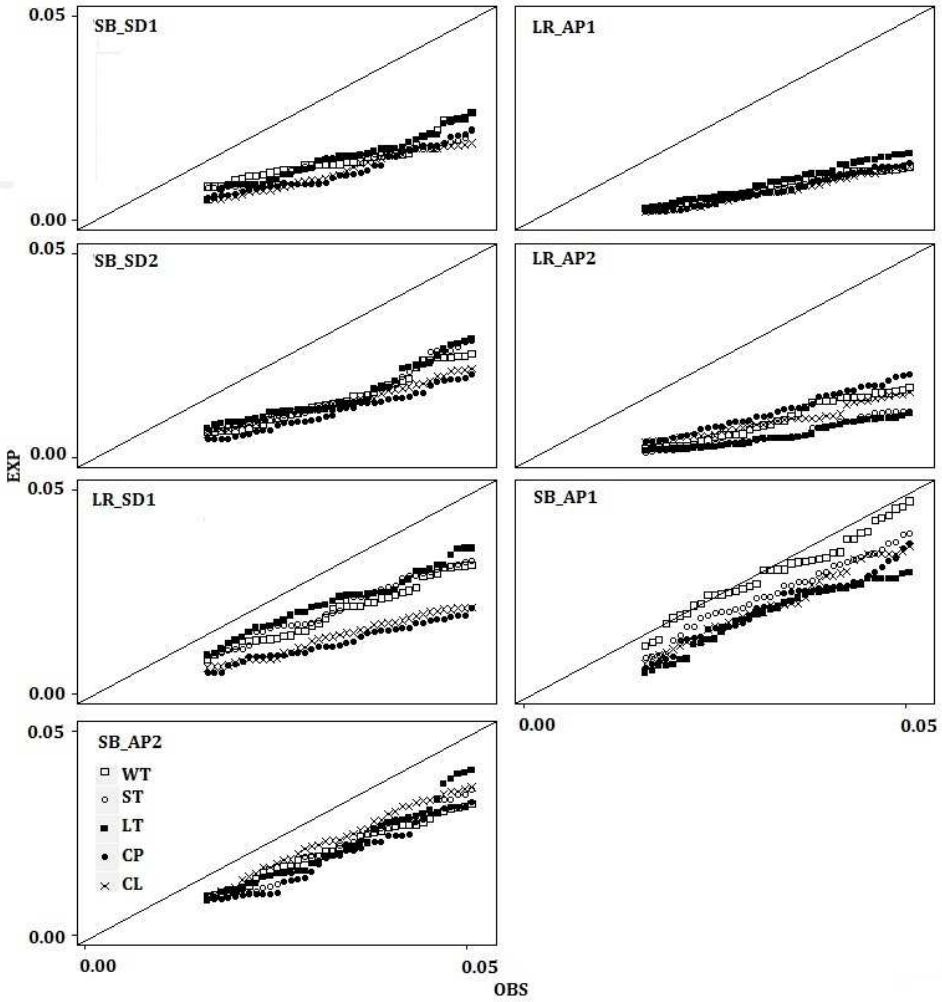

Figura 5: Q-Q plots entre los p-valores observados y esperados bajo la distribución uniforme estándar para cada método de GWAS comparado. SB_SD1 y SB_SD2: mancha borrosa en plántula (La Estanzuela, 2009), LR_SD1: roya de la hoja en plántula (La Estanzuela, 2009), SB_AP2: mancha borrosa en planta adulta (La Estanzuela, 2009-2010), LR_AP1: roya de la hoja en planta adulta (La Estanzuela, 2010), LR_AP2: roya de la hoja en planta adulta (EMAC, 2010) para los métodos, WT: regresión lineal simple sin transformación de los datos, ST: regresión lineal simple con transformación raíz cuadrada de los datos, LT: regresión lineal simple con transformación logaritmo de los datos, CP: regresión cumulativa probit y CL: regresión acumulativa logit

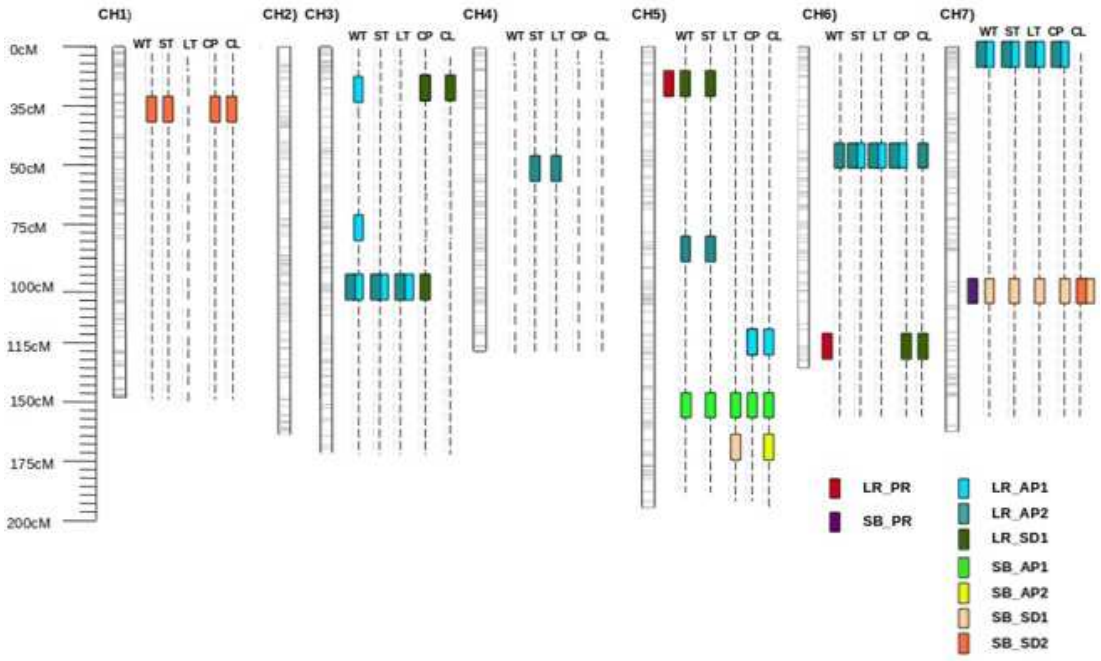

Figura 6: QTL detectados. SB_SD1 y SB_SD2: mancha borrosa en plántula (La Estanzula, 2009), LR_SD1: roya de la hoja en plántula (La Estanzuela, 2009), SB_AP2: mancha borrosa en planta adulta (La Estanzuela, 2009-2010), LR_AP1: roya de la hoja en planta adulta (La Estanzuela, 2010), LR_AP2: roya de la hoja en planta adulta (EMAC, 2010) para los métodos, WT: regresión lineal simple sin transformación de los datos, ST: regresión lineal simple con transformación raíz cuadrada de los datos, LT: regresión lineal simple con transformación logaritmo de los datos, CP: regresión cumulativa probit y CL: regresión acumulativa logit

3.5.2. Datos simulados

El comportamiento de la potencia y el FDR fue similar a través del número de categorías de fenotipado, para heredabilidades medias y altas (Fig. 7). Al variar el modelo genético subyacente, el patrón general se mantuvo para heredabilidades altas pero en general, no hubieron diferencias entre métodos, salvo con respecto al LT, que tuvo una menor potencia que el resto (Tabla 3).

El patrón general para el sesgo se mantuvo a través de los escenarios simulados. A través de los métodos de GWAS, tanto los métodos WT, ST y CP no resultaron significativamente diferentes entre sí. Los modelos LT y CL resultaron más significativamente diferentes del resto y entre sí.

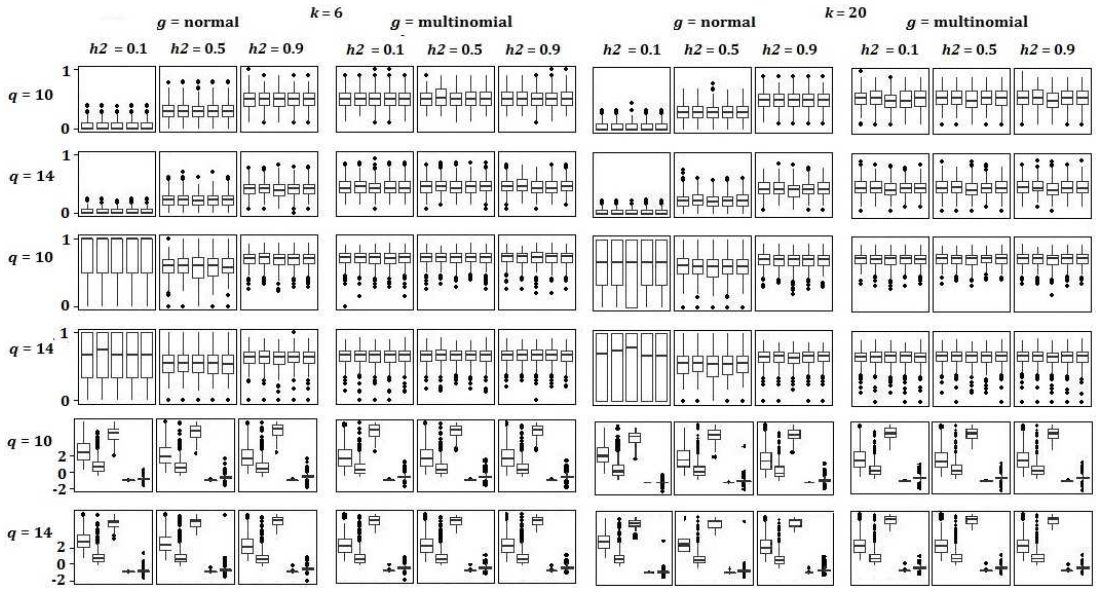

Figura 7: Potencia, FDR y sesgo en la detección de QTLs y estimación de sus efectos con números de QTL (*q*), heredabilidad (*h2*), número de categorías de fenotipado (*k*) y modelo genético subyacente para cinco métodos de GWAS: WT (regresión lineal simple sin transformación), ST (regresión lineal simple con transformación raíz cuadrada, LT (regresión lineal simple con transformación logaritmo, CP (regresión acumulativa con link probit) y CL (regresión acumulativa con link logit

Tabla 3: Potencia, FDR y sesgo en la detección de QTLs y estimación de sus efectos para WT (regresión lineal simple sin transformación), ST (regresión lineal simple con transformación raíz cuadrada, LT (regresión lineal simple con transformación logaritmo, CP (regresión acumulativa con link probit) y CL (regresión acumulativa con link logit)

modelo	potencia†	FDR¶	sesgo†
LT	0,38 [b]	0,50	6,49[a]
CL	0,40 [a]	0,50	2,34[b]
CP	0,40 [a]	0,50	-0,49[c]
WT	0,40 [a]	0,50	-0,64[c]
ST	0,40 [a]	0,50	-0,98[c]

†Diferencias significativas al p = 0,05 mediante la prueba de Friedman. Letras diferentes indican diferencias entre métodos obtenidas con la prueba de Nemenyi-Damico-Wolfe-Dunn

¶Sin diferencias significativas al p=0,05 para la prueba de Friedman

3.6. DISCUSIÓN

En este trabajo se compararon distintos métodos para realizar para el GWAS para la respuesta a enfermedades en cebada mediada en escala ordinal. Los métodos consistieron en dos variantes del modelo generalizado acumulativo (CP y CL, con función link probit y logit, respectivamente), dos transformaciones clásicas para el modelo lineal general (ST y LT, raíz cuadrada y logaritmo, respectivamente) y un modelo general sin transformación de la variable de respuesta (WT). Si bien tanto el uso de modelos generalizados como transformaciones se han usado en el mapeo de QTL para respuesta a enfermedades (Spyrides-Cunha et al., 2000; Castro et al. 2002; Roy et al. 2010; Wang et al. 2012) no queda claro cuál es la ventaja empírica de usar una una u otra.

En general, los métodos detectaron los mismos QTLs para la respuesta a cada enfermedad, de acuerdo con los resultados de los q-q plots. Esta similitud pudo deberse a un posible enmascaramiento de la naturaleza ordinal de los datos, al haber basado el fenotipado en los BLUEs obtenidos del análisis a campo. Sin embargo, el número de repeticiones verdaderas que se utilizó fue muy bajo como para que se alcance la distribución asintótica normal para las medias usadas. Además, las lecturas que conformaron las variables usadas fueron asignadas a categorías discretas según el juicio del fitopatólogo, de modo que tampoco está asegurado el cumplimiento del supuesto de normalidad, aunque estos valores se hayan tomado en una escala cuantitativa al aplicar los modelos generales.

De los QTLs hallados, tres mostraron una localización coincidente con reportes previos: *Rph2* (Borovkova et al. 1997) y *Rph11* (Feuerstein et al. 1990) para roya de la hoja en plántula y *Res-qtl-71412-30004* (Roy et al. 2011) para mancha borrosa en plántula (Fig. 3, Tabla sup. 1). Todos los métodos permitieron detectar QTL con localización coincidente a *Res-qtl-71412-30004*. Para *Rph2* y *Rph11*, no todos coincidieron: el QTL coincidente con *Rph2* fue detectado por WT y ST, mientras que el coincidente con *Rph11* fue detectado por CP y CL (Fig. 3, Tabla sup. 1). Para ambos QTLs, los métodos generales tuvieron una aparente mayor precisión

que los generalizados (Fig. 2). Si bien al comparar los Q-Q plots entre métodos estos sugieren una tasa de falsos positivos mayor para modelos generalizados que para generales, esto no es suficiente evidencia para cuestionar que la detección de *Rph11* sea un artefacto. En el caso de las simulaciones, el patrón observado fue el esperado según la teoría: mayor potencia y menores FDR y sesgo para valores altos de tamaño poblacional y heredabilidades (Beavis, 1998; Bernardo, 2004). Si bien al observar las gráficas aisladas para el FDR, este resultó mayor para heredabilidad alta que para media, la relación entre las medianas de la potencia y el FDR fue mayor a medida que la heredabilidad aumentó y el número de QTL disminuyó. De este modo, por cómo se definieron potencia y FDR, que involucró el número de verdaderos positivos, obtuvimos que el escenario óptimo es el previsto por la teoría: mejor comportamiento en la detección de QTL para menos QTLs y mayor herabilidad.

Las simulaciones fueron concebidas para generar distintos escenarios donde eventualmente los métodos de GWAS comparados difirieran (eg. variando el número de categorías fenotípicas y los modelos genéticos subyacentes). Si bien hubo diferencias entre los modelos genéticos para ambas métricas de precisión, estas no difirieron entre los distintos métodos de GWAS comparados, con la excepción de la potencia para el método LT, que fue menos potente que el resto. Tanto la potencia como el FDR fueron calculados en base al número de marcadores tomados como significativamente asociados al fenotipo o no, y esto a su vez depende del error por comparación múltiple usado. Si bien el umbral se definió teniendo en cuenta los grupos de pruebas independientes, es posible que algunas diferencias más sutiles entre métodos pasaran inadvertidas. No obstante, si las heredabilidades son altas, con una corrección por comparación múltiple exigente, seguiríamos detectando QTL de efecto mayor independientemente del método de GWAS usado (capítulo 1).

En general, para ambas enfermedades, los modelos se comportaron de la misma forma, coincidiendo con los resultados para la precisión a partir de las simulaciones. En los casos en que las diferencias fueron evidentes, el patrón siempre correspondió a la separación de los modelos generalizados, con curvas por encima de

los modelos generales. Si bien la teoría indica que cada modelo debe cumplir con las suposiciones adecuadas para ajustar correctamente los datos, en el caso donde los modelos difirieron, el modelo clásico sin transformar no tuvo peor comportamiento. Esto sugiere que la elección del modelo clásico de GWAS no afectaría negativamente a la detección de QTLs en el caso de variables ordinales.

Ambas enfermedad son ejemplos de infecciones fúngicas causadas por patógenos biotrófos (roya de la hoja) y necrotróficos (mancha borrosa) y la respuesta a la infección por cada uno supone bases genéticas bien distintas. Teniendo en cuenta el número de QTLs verdaderos y sus interacciones actuando sobre la expresión genotípica, según el modelo genético característico, se esperan que la distribución fenotípica y su vínculo con el genotipo sean diferentes para ambas enfermedades. No obstante, nuestros resultados sugieren que el ajuste de un modelo clásico de GWAS (ie. que supone normalidad de los errores) puede tener una performance igual o mejor que los otros métodos aquí comparados. En este contexto, el uso de un modelo clásico no invalidaría la inferencia de posición al mapear QTLs para la respuesta a ambos tipos de infección. Si bien se desconoce a priori el verdadero efecto de los QTLs mapeados para obtener una medida del sesgo en las estimaciones, los resultados de las simulaciones sugieren que la aproximación clásica también aportaría resultados insesgados.

La casi totalidad de los reportes de genes de resistencia a enfermedades de cebada realizados por GWAS han sido hechos sin considerar los problemas de ajustar modelos de errores normales a datos discretos. Estos resultados indican que, con el umbral adecuado para las pruebas de comparación múltiple, las diferencias entre modelos se hacen irrelevantes, por lo que los QTLs así reportados no serían artefactos.

3.7. BIBIOGRAFÍA

Abdurakhmonov I., Y. and A. Abdukarimov. 2008. Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources. International Journal of Plant Genomics 2008:1-18.

Bankina B. and Z. Gaile. 2009. Evaluation of barley disease development depending on varieties. Agronomy Research 7:198-203.

Beavis W. 1998. QTL analyses: Power, precision and accuracy. *In*: Molecular Dissection of Complex Traits. Paterson A. H. (ed). CRC Press. New York. pp.145-162.

Bernardo, R. 2004. What proportion of declared QTL in plants are false? Theoretical and Applied Genetics 109:419-424.

Bertrand C., Y. Collard and D. J. Mackill. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of The Royal Society 363:557-572.

Bonferroni, C.E. 1935. Il calcolo delle assicurazioni su gruppi di teste. *In*: Studi in Onore del Professore Salvatore Ortu Carboni. Roma. pp13-60.

Borovkova I. G., Y. Jin, B. J. Steffenson, A. Kilian, T. K. Blake and A. Kleinhofs. 1997. Identification and mapping of a rust resistance gene in barley line Q21861. Genome 40:236-241.

Castro A. J., X. Chen, P. M. Hayes, and S. J. Knapp. 2002. Coincident QTL which determine seedling and adult plant resistance to strip rust in barley. Crop Science 42:1701-1708.

Close T. J., P. R. Bhat, S. Lonardi, Y. Wu, N. Rostoks, L. Ramsay, A. Druka, N. Stein, J. T. Svensson, S. Wanamaker, S. Bozdag, M. L. Roose, M. J. Moscou, S. Chao, R. K. Varshney, P. Szücs, K. Sato, P. M. Hayes, D. E. Matthews, A. Kleinhofs, G. J. Muehlbauer, J. DeYoung, D. F. Marshall, K. Madishetty, R. D. Fenton, P. Condamine, A. Graner and R. Waugh. 2009. Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics 10:1-13.

Eckermann P. J., A. P. Verbyla, A. P. Cullis and B. R. Thompson. 2001. The analysis of quantitative traits in wheat mapping populations. Australian Journal of Agricultural Research 52:1195-1206.

Federer W. T. 1961. Augmented designs with one-way elimination of heterogeneity. Biometrics 17:447-473.

Feuerstein U., A. H. D. Brown, J. J.Burdon. 1990. Linkage of rust resistance genes from wild barley (*Hordeum spontaneum*) with isozyme markers. Plant Breeding 104:318-324.

Friedman M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the America Statistical Association. 32: 675-701.

Hayes, P., A Castro, L. Marquez-Cedillo, A. Corey, C. Henson, B.L. Jones, J.Kling, D. Mather, I. Matus, C. Rossi and K. Sato. 2003. Genetic diversity for quantitatively inherited agronomic and malting quality traits. In: R. Von Bothmer, Th. Van Hintum, H. Knupffer, and K.Sato (eds). Diversity in Barley (Hordeum vulgare).Elsevier Science. Amsterdam. pp:201-226.

HGCA. 2009. The barley disease management guide 2009 [En línea]. Consultado el 11 de enero de 2014. Disponible en: http://publications.hgca.com/publications/documents/cropresearch/BDMG_Complete2.pdf

Horsfall, J., R. Barratt. 1945. An improved grading system for measuring plant diseases. Phytopathology 35:655.

Jannink J. and B. Walsh. 2002. Association mapping in plant populations. *In*: Quantitative Genetics, Genomics and Plant Breeding. M. S. Kang (ed.). CABI International. Oxford. pp.59-68.

Kang H. M., N. Zaitlen, C. Wade, A. Kirby, D. Heckerman, M. J. Daly and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:709-1723.

Madden L.V. and G. Hughes. 1995. Plant disease incidence: distributions, heterogenity, and temporal analysis. Annual Review of Phytopathology 33:529-564.

Malosetti M., C. van der Linden, B. Vosman and F. van Eeuwijk. 2007. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics 175:879–889.

Nemenyi P.B. 1963. Distribution-free Multiple Comparisons [Tesis de doctorado]. Princeton: Princeton University. 254p.

Phipps R. H. and J.R. Park. 2002. Environmental benefits of genetically modified crops: Global and European perspectives on their hability to reduce pesticide use. Journal of Animal and Feed Sciences 11, 1-18.

Price A. L., N. Patterson And R. Plenge. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature 38:904-909.

Pritchard J. K., M. Stephens and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Roy J., K. Smith, G. Muehlbauer, S. Chao, T. Close and B. Steffenson. 2010. Association mapping of spot blotch resistance in wild barley. Molecular Breeding 26:243-256.

Spyrides-Cunha M., C. Demétrio and L. Camargo. 2000. Proportional odds model applied to mapping of disease resistance genes in plants. Genetics and Molecular Biology 23: 223-227.

Stakman E. C., D. M. Stewart and W. Q. Loegering. 1962. Identification of physiologic races of *Puccinia graminis* var tritici. U.S. Department of Agriculture. ARS-E 6/7. 53 pp.

Steffenson B., P. Hayes and A. Kleinhofs. 1996. Genetics of seedling and adult plant resistance to net blotch (*Pyrenophora teresf.* Teres) and spot blotch (*Cochliobolus sativus*) in barley. Theoretical and Applied Genetics 92:552-558.

Verbyla A. P., P. J. Eckermann, R. Thompson and B. R. Cullis. 2003. The analysis of quantitative trait loci in multi-environment trials using a multiplicative mixed model. Australian Journal of Agricultural Research 54:1395-1408.

Wang, M., N. Jiang, T. Jia, L. Leach, J. Cockram, J. Comadran, P. Shaw R. Waugh, L. Ramsay and B. Thomas. 2012. Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theoretical Applied Genetics 124:233–246.

Xu S. and W.R. Atchey. 1996. Mapping quantitative trait loci for complex binary disease using line crosses. Genetics 143:1417-1424.

Yu J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovic and E. S. Buckler. 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38:203-208.

Yang R., N. Yi and S. Xu. 2006. Box-Cox transformation for QTL mapping. Genetica 128:133-143.

Zadoks J. C., T. T. Chang, and C. F. Konzak. 1974. A decimal code for the growth stages of cereals. Weed Research 14:415-421.

Zhao K., M. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, and P. Marjoram. 2007. An Arabidopsis example of association mapping in structured samples. PLoS Genetics 3:71-82.

4. DISCUSIÓN GENERAL

El propósito principal de este trabajo fue comparar distintos métodos para el GWAS de variables fenotípicas ordinales. Este tipo de variables, en particular la respuesta a enfermedades, tienen un comportamiento estadístico diferente al supuesto por los modelos convencionales para llevar a cabo el GWAS. Si bien existen distintos métodos estadísticos para este tipo de variables, no existen trabajos previos que den cuenta de la eficacia general o específica para un estudio determinado. Tener una medida de la eficacia de cada método podría potenciar el mejoramiento genético para incorporar fuentes de resistencia en el mejoramiento genético o para el diagnóstico de enfermedades.

Los métodos de GWAS para variables ordinales comparados consistieron en dos tipos de transformaciones clásicas y dos tipos de modelos lineales generalizados, contra un modelo lineal general de GWAS. Esto se implementó mediante dos enfoques alternativos: un enfoque de simulaciones de distintos escenarios poblacionales (simulando tanto datos genéticos como fenotípicos) y uno mixto, de simulaciones y datos reales, provenientes de ensayos para la respuesta a las infecciones por mancha borrosa y roya de la hoja en cebada de programas de mejoramiento de América Latina y del ICARDA/CIMMYT (México). Ambos enfoques supusieron la utilización de metodología previa junto con el desarrollo de nuevas aproximaciones, como la utilización de las pruebas no paramétricas de Friedman y Nemenyi-Damico-Wolfe-Dunn sobre las medianas de distintas métricas de eficacia para llevar a cabo las comparaciones entre los métodos de GWAS.

Para ambos enfoques en los que se dividió el trabajo se obtuvieron 4 resultados principales: 1) que el comportamiento de los métodos comparado depende del número de QTL, el tamaño poblacional y las heredabilidades según los esperado por la teoría, 2) que el número de categorías fenotípicas tiene un impacto irrelevante sobre este comportamiento, 3) que el modelo genético afecta el comportamiento de los modelos para heredabilidades bajas y 4) que el modelo general sin transformar para GWAS de variables ordinales se comportó de manera equivalente a los demás

métodos. Estos resultados sugieren que el procedimiento algorítmico para comparar los métodos de GWAS mediante simulaciones es una manera efectiva para reproducir los patrones esperados por la teoría, tanto las escalas en que se miden las variables como los modelos genéticos subyacentes tienen un efecto irrelevante sobre la performance relativa de los modelos, para heredabilidades medias a altas y por último, que los desvíos de la normalidad no implicarían la invalidez del reporte de QTL mediante modelos clásicos. Trabajos subsiguientes deberían tener en cuenta la comparación de otras alternativas no exploradas para el GWAS de variables ordinales, como son los modelos bayesianos y no-paramétricos, además de distintos tipos de corrección por comparación múltiple y estructura poblacional. Esto podría revelar diferencias más sutiles entre los métodos, dependiendo del contexto en el cual se realiza el estudio (ie. los recursos con los que se cuenta y la arquitectura genética).
Nuestro resultados indican que con un umbral para las pruebas de comparación múltiple lo suficientemente exigente, los modelos clásicos de errores normales pueden captar los QTLs de efecto mayor, relevantes para el mejoramiento. Esto implicaría que la mayoría de los reportes de QTLs para resistencia a enfermedades (mediante ajuste de modelos clásicos) no se corresponderían a artefactos, haciendo que la combinación de modelo de GWAS clásico y umbral exigente pueda seguir siendo un procedimiento todavía válido para la detección de factores de resistencia.

5. BIBLIOGRAFÍA

Abdolkarim M., M. Torabi, S. Rezaee And F. Afshari. 2011. Virulence genes and pathotypes of *Puccinia hordei* Otth causing leaf rust on barley in some areas of Iran. Seed and Plant Improvement Journal 27:89-102.

Abdurakhmonov I., Y. And A. Abdukarimov. 2008. Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources. International Journal of Plant Genomics 2008:1-18.

Bankina B. And Z. Gaile. 2009. Evaluation of barley disease development depending on varieties. Agronomy Research 7:198-203.

Beavis W. 1998. QTL analyses: Power, precision and accuracy. *In*: Molecular Dissection of Complex Traits. Paterson A. H. (ed). CRC Press. New York. pp.145-162.

Bernardo, R. 2004. What proportion of declared QTL in plants are false? Theoretical and Applied Genetics 109:419-424.

Bertrand C., Y. Collard And D. J. Mackill. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of The Royal Society 363:557-572.

Bonferroni, C.E. 1935. Il calcolo delle assicurazioni su gruppi di teste. *In*: Studi in Onore del Professore Salvatore Ortu Carboni. Roma. Pp13-60.

Borovkova I. G., Y. Jin, B. J. Steffenson, A. Kilian, T. K. Blake And A. Kleinhofs. 1997. Identification and mapping of a rust resistance gene in barley line Q21861. Genome 40:236-241.

Casella G. And R. Berger. 1990. Statistical inference. Second Edition. Duxbury. California. 688p.

Castro A. J., X. Chen, P. M. Hayes, and S. J. Knapp. 2002. Coincident QTL which determine seedling and adult plant resistance to strip rust in barley. Crop Science 42:1701-1708.

Close T. J., P. R. Bhat, S. Lonardi, Y. Wu, N. Rostoks, L. Ramsay, A. Druka, N. Stein, J. T. Svensson, S. Wanamaker, S. Bozdag, M. L. Roose, M. J.

Moscou, S. Chao, R. K. Varshney, P. Szücs, K. Sato, P. M. Hayes, D. E. Matthews, A. Kleinhofs, G. J. Muehlbauer, J. Deyoung, D. F. Marshall, K. Madishetty, R. D. Fenton, P. Condamine, A. Graner And R. Waugh. 2009. Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics 10:1-13.

Christensen R. H. B. 2012. Ordinal Regression Models for Ordinal Data. R package versiÓn 2012. 01-19

Coppieters W., A. Kvasz, F. Farnir, J. J. Arranz, B. Grisart, M. Mackinnon And M. Georges. 1998. A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sib pedigrees: application to milk production in a granddaughter design. Genetics 149:547-55.

Diao G. And D. Y. Lin. 2006. Improving the power of association tests for quantitative traits in family studies. Genetic Epidemiology 30:301-313.

Eckermann P. J., A. P. Verbyla, A. P. Cullis And B. R. Thompson. 2001. The analysis of quantitative traits in wheat mapping populations. Australian Journal of Agricultural Research 52:1195-1206.Falconer D. S. and T. F. C. Mackay .1996. Introduction to Quantitative Genetics. Edinburgh: Pearson. 459p.

Federer W. T. 1961. Augmented designs with one-way elimination of heterogeneity. Biometrics 17:447-473.

Feuerstein U., A. H. D. Brown, J. J.Burdon. 1990. Linkage of rust resistance genes from wild barley (*Hordeum spontaneum*) with isozyme markers. Plant Breeding 104:318-324.

Friedman M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the America Statistical Association. 32: 675-701.

Hackett C. And J. Weller. 1995. Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics 51:1252-1263.

Harville D. And R. Mee. 1984. A mixed-model procedure for analyzing ordered categorical data. Biometrics 40:393-408.

Hayes, P., A Castro, L. Marquez-Cedillo, A. Corey, C. Henson, B.L. Jones, J.Kling, D. Mather, I. Matus, C. Rossi and K. Sato. 2003. Genetic diversity for quantitatively inherited agronomic and malting quality traits. In: R. Von Bothmer, Th. Van Hintum, H. Knupffer, and K.Sato (eds). Diversity in Barley (Hordeum vulgare).Elsevier Science. Amsterdam. pp:201-226.

Henderson C. R.1984. Applications of Linear Models in Animal Breeding. L.R. Schaeffer (ed). Third edition. University of Guelph. Guelph.

Hennig, C. 2013. Fpc: Flexible procedures for clustering. R package version 2.1-5.

HGCA. 2009. The barley disease management guide 2009 [En línea]. Consultado el 11 de enero de 2014. Disponible en:http://publications.hgca.com/publications/documents/cropresearch/BDMG_Complete2.pdf.

Iwata, H., K. Ebana And S. Fukuoka. 2009. Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among *Oryza sativa L.* germplasms. Theoretical and Applied Genetics 18:865-880.

Jannink J. And B. Walsh. 2002. Association mapping in plant populations. *In*: Quantitative Genetics, Genomics and Plant Breeding. M. S. Kang (ed.). CABI International. Oxford. pp.59-68.

Kang H. M., N. Zaitlen, C. Wade, A. Kirby, D. Heckerman, M. J. Daly And E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:709-1723.

Madden L.V. And G. Hughes. 1995. Plant disease incidence: distributions, heterogenity, and temporal analysis. Annual Review of Phytopathology 33:529-564.

Malosetti M., C. Van Der Linden, B. Vosman And F. Van Eeuwijk. 2007. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics 175:879–889.

Nemenyi P.B. 1963. Distribution-free Multiple Comparisons [Tesis de doctorado]. Princeton: Princeton University. 254p.

Pearson T. A. And T.A Manolio. 2008. How to interpret a genome-wide association study. The Journal of the American Medical Association 299: 1335-1344.

Peterson R. F., A.B Campbell And A.E. Hannah. 1948 A Diagrammatic Scale for Estimating Rust Intensity of Leaves and Stem of Cereals. Canadian Journal of Research Section 26, 496-500.

Phipps R. H. and J.R. Park. 2002. Environmental benefits of genetically modified crops: Global and European perspectives on their hability to reduce pesticide use. Journal of Animal and Feed Sciences 11, 1-18.

Price A. L., N. Patterson And R. Plenge. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature 38:904-909.

Pritchard J. K. And M. Przeworski. 2001. Linkage disequilibrium in humans: models and data. American Journal of Human Genetics 69:1-14.

Pritchard J. K., M. Stephens And P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

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

Roy J., K. Smith, G. Muehlbauer, S. Chao, T. Close And B. Steffenson. 2010. Association mapping of spot blotch resistance in wild barley. Molecular Breeding 26:243-256.

Spyrides-Cunha M., C. Demétrio And L. Camargo. 2000. Proportional odds model applied to mapping of disease resistance genes in plants. Genetics and Molecular Biology 23: 223-227.

Stakman E. C., D. M. Stewart And W. Q. Loegering. 1962. Identification of physiologic races of *Puccinia graminis* var tritici. U.S. Department of Agriculture. ARS-E 6/7. 53 pp.

Steffenson B., P. Hayes And A. Kleinhofs. 1996. Genetics of seedling and adult plant resistance to net blotch (*Pyrenophora teresf.* Teres) and spot blotch (*Cochliobolus sativus*) in barley. Theoretical and Applied Genetics 92:552-558.

Verbyla A. P., P. J. Eckermann, R. Thompson And B. R. Cullis. 2003. The analysis of quantitative trait loci in multi-environment trials using a multiplicative mixed model. Australian Journal of Agricultural Research 54:1395-1408.

Wang, M., N. Jiang, T. Jia, L. Leach, J. Cockram, J. Comadran, P. Shaw R. Waugh, L. Ramsay And B. Thomas. 2012. Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theoretical Applied Genetics 124:233–246.

Wu R., C. Ma And G. Casella. 2010. Statistical Genetics of Quantitative Traits: Linkage, Maps and QTL. Springer. New York. 368p.

Xu S. And W.R. Atchey. 1996. Mapping quantitative trait loci for complex binary disease using line crosses. Genetics 143:1417-1424.

Xu S. 2003. Theoretical basis of the Beavis effect. Genetics 165:2259-2268.

Yu J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. Mcmullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovic And E. S. Buckler. 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38:203-208.

Yang R., N. Yi And S. Xu. 2006. Box-Cox transformation for QTL mapping. Genetica 128:133-143.

Zhao K., M. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, And P. Marjoram. 2007. An Arabidopsis example of association mapping in structured samples. PLoS Genetics 3:71-82.