Variance heterogeneity genome-wide mapping for cadmium in bread wheat reveals novel genomic loci and epistatic interactions

Waseem Hussain[1*], Malachy Campbell[2], Diego Jarquin[1], Harkamal Walia[1], and Gota Morota[2*]

[1]Department of Agronomy and Horticulture, University of Nebraska-Lincoln
[2]Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University

Keywords: cadmium, epistasis, genome-wide association analysis, variance heterogeneity, wheat.

Running title: vGWAS for wheat grain cadmium

ORCID: 0000-0002-6861-0193 (WH), 0000-0002-8257-3595 (MTC), 0000-0002-5098-2060 (DJ), 0000-0002-9712-5824 (HW), and 0000-0002-3567-6911 (GM).

* Corresponding author:

Waseem Hussain
Department of Agronomy and Horticulture
University of Nebraska-Lincoln
Lincoln, Nebraska 68583 USA.
E-mail: waseem.hussain@unl.edu

Gota Morota
Department of Animal and Poultry Sciences
Virginia Polytechnic Institute and State University
175 West Campus Drive
Blacksburg, Virginia 24061 USA.
E-mail: morota@vt.edu

Abstract

Genome-wide association mapping identifies quantitative trait loci (QTL) that influence the mean differences between the marker genotypes for a given trait. While most loci influence the mean value of a trait, certain loci, known as variance heterogeneity QTL (vQTL) determine the variability of the trait instead of the mean trait value (mQTL). Identification of genetic variants that affect variance heterogeneity can provide insights into the biological mechanisms that control variation, phenotypic plasticity, and epistasis. In the present study, we performed variance heterogeneity genome-wide association studies (vGWAS) for grain cadmium (Cd) concentration using a hard-red winter wheat (*Triticum aestivum L.*) association mapping panel. We used double generalized linear model (DGLM) and hierarchical generalized linear model (HGLM) to identify vQTL associated with grain Cd. We identified novel vQTL regions on chromosomes 2A and 2B that contribute to the Cd variation and loci that affect both mean and variance heterogeneity (mvQTL) on chromosome 5A. In addition, our results demonstrated the presence of epistatic interactions between vQTL and between vQTL and mvQTL, which could explain variance heterogeneity. Several candidate genes that were associated with the regulation of mineral content in plants were identified; these included genes encoding a homeobox-leucine zipper family protein, ABC transporter, MADS-box transcription factor, plant peroxidase, and glycosyltransferase. Overall, we provide novel insights into the genetic architecture of grain Cd concentration and report the first application of vGWAS in wheat. Moreover, our findings indicated that epistasis is an important mechanism underlying natural variation for grain Cd concentration.

Background

Genome-wide association studies (GWAS) are routinely conducted to study the genetic basis of important traits in crops. GWAS use populations of related individuals and link phenotypic variation with dense genetic marker data using a linear modeling framework (Xiao et al., 2017). Standard GWAS approaches seek to identify trait-marker associations that influence the mean phenotypic values. However, differences in the variance between genotypes are also under genetic control (Shen et al., 2012). As a result, several recent studies have identified loci associated with differences in variance between genotypes (Corty and Valdar, 2018; Corty et al., 2018; Cao et al., 2014). Such genetic variants that affect the variance heterogeneity of traits have been referred to as variance heterogeneity quantitative trait loci (vQTL).

Variance heterogeneity-based genome-wide association studies (vGWAS) have emerged as a new approach for identifying and mapping vQTL. vQTL contribute to variability, which is undetected through standard statistical mapping (bi-parental or association) procedures (Forsberg and Carlborg, 2017; Rönnegård and Valdar, 2011; Shen et al., 2012). It has been argued that variance heterogeneity between genotypes can be partially explained by epistasis or gene-by-environment interactions (Brown et al., 2014; Forsberg and Carlborg, 2017; Young et al., 2018). Thus, vQTL can provide insights into epistasis or phenotypic plasticity (Young et al., 2018; Nelson et al., 2013). Moreover, these vGWAS frameworks can serve as tractable approaches to reduce the search space when assessing epistasis among markers (Brown et al., 2014; Wei et al., 2016).

Numerous studies have reported vQTL associated with diverse phenotypes, including the tendency to left-right turning and bristles (Mackay and Lyman, 2005) and locomotor handedness (Ayroles et al., 2015) in *Drosophila*; coat color (Nachman et al., 2003), circadian activity, and exploratory behavior (Corty et al., 2018) in mice; thermotolerance (Queitsch et al., 2002), flowering time (Salom et al., 2011), and molybdenum concentration (Forsberg et al., 2015; Shen et al., 2012) in *Arabidopsis*; litter size in swine (Sell-Kubiak et al., 2015);

urinary calcium excretion in rats (Perry et al., 2012); and body mass index (Yang et al., 2012; Young et al., 2018), sero-negative rheumatoid arthritis (Wei et al., 2017), psoriasis (Wei et al., 2018), and serum urate (Topless et al., 2015) in humans. In plants, vGWAS have been limited to few species, including *Arabidopsis* (Forsberg et al., 2015; Shen et al., 2012) and maize (Kusmec et al., 2017). To date, vGWAS have been very limited.

Methodologically, vQTL have been detected by performing statistical tests searching for unequal variance for a quantitative trait between the marker genotypes (Dumitrascu et al., 2018). The most common statistical tests used to identify vQTL include Levene's test (Par et al., 2010), Brown-Forysthe test (Brown and Forsythe, 1974), squared residual value linear modeling (Struchalin et al., 2012), and correlation least squares test (Brown et al., 2014). However, these methods have certain drawbacks when applied to genetic data. For example, Levene's and Brown-Forsythe tests are sensitive to deviations from normality and have an inherent inability to model continuous covariates (Rönnegård and Valdar, 2012; Dumitrascu et al., 2018).

Double generalized linear model (DGLM) has emerged as an alternative approach to model the variance heterogeneity for genetic studies (Rönnegård and Valdar, 2011). In DGLM, sample means and residuals are modelled jointly. Here, generalized linear models (GLM) are fitted by including only the fixed effects in the linear predictor(s) for the mean and dispersion. It is important to correct for population structure, which can otherwise lead to spurious associations in GWAS (Patterson et al., 2006). In DGLM, population structure can be corrected by incorporating the first few principal components of a genomic relationship matrix (GRM) (Patterson et al., 2006) as fixed covariates in the model. However, the first few principal components may not be sufficient to account for complex population structure or family relatedness (Hoffman, 2013; Sul et al., 2018). Alternatively, we can fit linear mixed models (LMM) to strictly correct for population structure, where the whole GRM can be modeled as random effects. Hierarchical generalized linear model (HGLM) has been proposed as an extension of the DGLM to model random effects in the mean component (Rönnegård and Valdar, 2012; Tan et al., 2014). In HGLM, the GRM can be used to model correlated random effects and account for population structure.

We applied a vGWAS framework to examine the genetic architecture of Cd accumulation in wheat grains in the current study. Cd is a heavy metal that is highly toxic to human health (Menke et al., 2008). Identifying genetic variants that control low-grain Cd concentration in wheat is necessary to understand the basis for phenotypic variation in grain Cd and can help accelerate the development of low Cd wheat varieties. A recent study assessed natural variation in grain Cd in bread wheat by conducting GWAS (Guttieri et al., 2015). However, only a fraction of phenotypic variation could be explained by the top marker associations, indicating that grain Cd concentration is a complex trait that is influenced by multiple loci and/or loci with non-additive effects (Guttieri et al., 2015). Given the genetic complexity of Cd in wheat, we hypothesized that variation in grain Cd concentration in wheat is influenced by vQTL that are likely to be involved in epistatic interactions; this would allow us to capture additional variation that are not accounted for in a standard GWAS approach.

In this study, we sought to provide additional insights into natural variation in grain Cd concentration in bread wheat through vGWAS using a publicly available hard-red winter wheat association mapping panel (`https://triticeaetoolbox.org/wheat/`). To achieve this, we used DGLM and HGLM to perform vGWAS. Previously, Guttieri et al. (2015) conducted standard GWAS using this association panel and identified a single mean effect QTL (mQTL) for grain Cd concentration on chromosome 5A. In addition, we aimed to understand the basis of vQTL by searching for pairwise epistatic interactions among vQTL and mQTL and add biological context to the identified vQTL regions by unraveling candidate genes within these genomic intervals. To our knowledge, the present study is the first to conduct vGWAS and identify vQTL associated with grain Cd concentration in wheat.

Materials and Methods

Plant materials and genotyping

We analyzed a publicly available dataset comprising of phenotypes for grain mineral concentration for $n = 299$ genotyped hard-red winter wheat accessions. The details of the study are discussed in Guttieri et al. (2015), and access to the data is available at `http://triticeaetoolbox.org/wheat/`. Here, we focused on grain Cd concentration (mg/kg) averaged across two years in one location (Oklahoma, USA). We combine the data across years due to non-significant genotype x year interactionGuttieri et al. (2015). The association panel was genotyped using a 90K iSelect Infinium array (Wang et al., 2014b). We used a filtered marker data set consisting of single nucleotide polymorphism (SNP) markers from the 90K iSelect Infinium array as described by Guttieri et al. (2015). All the SNP markers were physically anchored on the new reference genome of hexaploid wheat RefSeq v1.0 (Appels et al., 2018).

Statistical modeling

We used DGLM and HGLM to detect VQTL in the current study. The description of models used is given below.

DGLM

DGLM is a parametric approach that can be used to jointly model the mean and dispersion using a GLM framework (Smyth, 1989). The DGLM model works iteratively by first fitting a linear model to estimate the mean effects (mQTL). The squared residuals are used to estimate the dispersion effects (vQTL) using GLM with a gamma-distributed response and the log link function. This process is cycled until convergence. Here, we extended the DGLM model to marker-based association analysis according to Rönnegård and Valdar (2011). The mean part of DGLM was as follows:

$$\mathbf{y} = \mathbf{1}\mu_m + \mathbf{X}\boldsymbol{\beta} + \mathbf{S}_j a_{mj} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is the Cd concentration (mg/kg); $\mathbf{1}$ is the column vector of 1; μ_m is the intercept; $\mathbf{X}$ is $n \times 4$ covariate matrix of the top four principle components (PCs) obtained by performing principal component analysis (PCA) of marker data using the SNPRelate R package (Zheng et al., 2012); $\boldsymbol{\beta}$ is the regression coefficients for the covariates; $\mathbf{S}_j$ $\in$(0,2) is the vector containing the number of reference allele at the marker j, a_{mj} is the effect size or allele substitution effect of the jth marker; and $\boldsymbol{\epsilon}$ is the residual. We assumed

$$\epsilon \sim N(0, \mathbf{I}\sigma^2_\epsilon)$$

$$log(\sigma^2_\epsilon) = \mathbf{1}\mu_v + \mathbf{S}_j a_{vj},$$

where $\mathbf{I}$ is the identity matrix; σ^2_ϵ is the residual variance; and $\mathbf{1}\mu_v$ and a_v are the intercept and marker regression coefficients for the variance part of the model, respectively. While we fit separate effects for the mean using a standard linear model and for the variance using the squared residuals in gamma distributed GLM with a log link function, this is equivalent to modeling $\mathbf{y} \sim N(\mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{S}a_{mj}, \exp(\mathbf{1}\mu_v + \mathbf{S}_j a_{vj})$ or $\epsilon \sim N(0, \exp(\mathbf{1}\mu_v + \mathbf{S}_j a_{vj}))$ in equation (1).

The DGLM model was fitted using the dglm package (`https://cran.r-project.org/web/packages/dglm/index.html`) in R statistical computing environment (R Core Team, 2018). SNP markers were fitted one by one, and for each marker, the effect sizes, standard errors, and p-values were obtained for the mean and dispersion components. To account for multiple testing, we determined the effective number of independent tests (Meff) using the method described by Li and Ji (2005). Subsequently, a genome-wide significance threshold level ($P < 1.44 \times 10^{-5}$) was determined using the following formula:

$$\alpha_p = 1 - (1 - \alpha_e)^{\frac{1}{\mathrm{Meff}}},$$

where α_p is the genome-wide significance threshold level, and α_e is the desired level of significance (0.05).

HGLM

One approach to correct for population structure is to perform PCA of the marker matrix, extract the first few principal components, and fit them as covariates to correct for population structure, as in the DGLM approach. However, this approach captures some but not all population structure (Hoffman, 2013). To explicitly account for population structure and kinship in GWAS, LMM have been proposed as alternative methods that allow the genetic relationships between individuals to be modeled as random effects. To perform vGWAS in the LMM framework and to identify genome-wide vQTL, we used a HGLM approach. HGLM (Lee and Nelder, 1996) is a class of GLM and is a direct extension of the DGLM that allows joint modelling of the mean and dispersion parts and introduces random effects as a linear predictor for the mean (Lee et al., 2006; Rönnegård and Carlborg, 2007). The mean part of HGLM was given as follows:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{S}_j a_{mj} + \mathbf{Z}\mathbf{u} + \epsilon,$$

assuming that

$$u \sim N(0, \mathbf{G}\sigma_u^2),$$

where $\mathbf{Z}$ is the incident matrix of random effects; u is the vector of random effects with $\mathrm{Var}(u) = \mathbf{G}{\sigma^2}_u$; $\mathbf{G}$ is the GRM of VanRaden (2008); and σ_u^2 is the additive genetic variance. A log link function is used for the residual variance given by $\exp(\mathbf{S}_j, a_{vj})$, which is equivalent to modeling $\mathbf{y}|a_{mj}, \mathbf{u}, a_{vj} \sim N(\mathbf{S}_j a_{mj}, \mathbf{Zu}, \exp(\mathbf{S}_j, a_{vj}))$.

We fitted HGLM using the hglm R package (Rönnegård et al., 2010). We reformulated the term $\mathbf{Zu}$ as $\mathbf{Z}^*\mathbf{u}^*$, where $\mathbf{u}^* \sim N(0, \mathbf{I}\sigma_u^2)$; $\mathbf{Z}^* = \mathbf{Z}_0\mathbf{L}$; $\mathbf{L}$ is the Cholesky factorization of the $\mathbf{G}$ matrix; and $\mathbf{Z}_0$ is the identity matrix (Rönnegård and Carlborg, 2007). Markers treated as fixed effects were fitted one by one, and for each marker, the effect sizes, standard errors, and p-values were obtained for the mean and dispersion components. The genome-wide significance threshold level was derived as described in the DGLM analysis.

Epistasis analysis

We investigated the extent of epistasis that was manifested through variance heterogeneity. All the possible pairwise interaction analyses for markers that were associated with grain Cd concentration were performed using the following two markers at a time epistatic model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{S}_j a_j + \mathbf{S}_k a_k + (\mathbf{S}_j\mathbf{S}_k)v_{jk} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is the vector of Cd concentration (mg/kg); $\mathbf{X}$ is the incident matrix for the first four PCs; $\boldsymbol{\beta}$ is the regression coefficients for the PCs; $\mathbf{S}_j$ and $\mathbf{S}_k$ are SNP codes for the jth and kth markers, respectively; a_j and a_k are the additive effects of the markers j and k, respectively; and v_{jk} is the additive × additive epistatic effect of the jth and kth marker. We used Bonferroni correction to account for the multiple testing.

Candidate gene identification

We performed candidate gene identification for the SNP markers associated with variance heterogeneity. We used the Ensembl Plants browser (Bolser et al., 2017) to retrieve the candidate genes and functional annotations (`http://plants.ensembl.org/Triticum_aestivum/Info/Index`) and the International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v1.0 annotations (Appels et al., 2018) available at `https://wheat-urgi.versailles.`

inra.fr/Seq-Repository/Annotations. For candidate gene analysis, we first determined the positions of significant SNP markers, and the interval was defined as the distance between the lowest and highest markers based on the position of SNPs. For example, if the position of the lowest SNP and highest SNP was 715,333,165 bp and 717,146,211 bp in the vQTL region on chromosome 2A, we defined 2A as the 715,333,165-717,146,211 interval for candidate gene identification. After defining the interval for the 2A (2A: 715,333,165-717,146,211) and 2B (2B: 691,780,716- 701,097,263 bp) regions, we explored the intervals using the Ensembl Plants browser and extracted the Gene IDs within these intervals. The Gene IDs within the defined interval on chromosomes 2A and 2B were analyzed using the IWGSC RefSeq v.1.0 (Appels et al., 2018) integrated genome annotations to obtain the predicted genes and functional annotations.

Data availability

The wheat phenotypic and genotypic data can be downloaded from http://triticeaetoolbox.org/wheat/ and also available on the GitHub repository https://github.com/whussain2/vGWAS. The R code used for the analysis is available on the GitHub repository https://github.com/whussain2/vGWAS. File S1 contains Supplementary Table S1 and Figures S1-S4. File S2 contains a list of all candidate genes and annotations associated with the vQTL on chromosomes 2A and 2B.

Results

Variance heterogeneity GWAS provide additional insights into natural variation in grain Cd

Although grain Cd concentration is a highly heritable trait, recent GWAS revealed that significant loci can only explain a fraction of the variation for this trait (Guttieri et al., 2015). Thus, to further examine natural variation for grain Cd concentrations in wheat, we performed vGWAS using genotypic and phenotypic data for 299 diverse hard-red winter wheat accessions (Guttieri et al., 2015). The DGLM and HGLM approaches were used to detect vQTL while controlling for population structure.

First, we conducted the DGLM-based analysis to each SNP and calculated the P-values for mean and dispersion effects. We classified the QTL into the following categories: mQTL, which contributes to difference in the means between marker genotypes; vQTL, which influences the variability between the genotypes; and mean-variance QTL (mvQTL), which contributes to differences in both the mean and variance between the genotypes.

Based on the DGLM, we identified two vQTL associated with the variance heterogeneity of Cd concentration. One vQTL on 2A contained four SNP markers, and one vQTL on 2B contained 17 SNP markers (Figure 1 and Supplementary File S1: Table S1). The four SNP markers associated with the vQTL region on the chromosome 2A region spanned the physical distance of 1.81 Mb; all SNP markers were located within the 0 kb linkage disequilibrium (LD) block (Supplementary File S1: Figure S1). The vQTL region on 2B associated with 17 SNP markers spanned the physical distance of 9.32 Mb, and the SNP markers were located within four LD blocks of sizes 0, 1, 1, and 204 kb (Supplementary File S1: Figure S2).

In addition, we identified a single mvQTL (containing four SNP markers) associated with both mean and variance heterogeneity on chromosome 5A (Figure 1 and Table S1). The markers associated with mvQTL on chromosome 5A were identical to those obtained in the original GWAS analysis according to Guttieri et al. (2015), indicating that this region affects both the mean and the variance heterogeneity (Supplementary File S1: Figure S1). Moreover, these results showed that DGLM serves as an accurate framework to jointly detect mean and variance QTL and provides additional insights into phenotypic variation that would otherwise not be captured by standard GWAS.

Considering that population stratification was detected using the association panel used in this study, we next used HGLM, which captures population substructure between individuals using the **G** matrix. This model extends the DGLM framework and allows a random effect to fit the mean regression component. vGWAS based on HGLM revealed the same results as those obtained using DGLM and showed identical vQTL on chromosomes 2A and 2B and mvQTL on chromosome 5A associated with variance heterogeneity of Cd concentration.

Variance heterogeneity loci can be partially explained by epistasis

Although the interpretation of vQTL results remains controversial and is dependent on the experimental design and the parameterization of the mean component of the model, one possible explanation for the vQTL is the presence of epistatic interactions between marker genotypes (Forsberg and Carlborg, 2017). Thus, we next sought to investigate whether the vQTL identified in this study are involved in epistatic interactions. We investigated all significant markers (25 markers) associated with mvQTL on chromosome 5A and vQTL on chromosomes 2A and 2B and explored all possible pairwise additive × additive epistatic interactions. Interestingly, we detected significant additive × additive interactions between the markers (Figure 2). The interaction was more evident between mvQTL on chromosome 5A and vQTL on chromosomes 2A and 2B. Specifically, all the markers associated with the 5A mvQTL region revealed highly significant interactions with all the markers associated with the 2A and 2B vQTL regions. Interactions between vQTL on 2A and 2B chromosomes were also observed; however, the interactions were less evident, and only a few markers within these regions showed statistically significant interactions. Taken together, these results suggested that the vQTL and mvQTL may be manifested because of pairwise epistatic interactions.

Candidate gene identification

We investigated the biological basis of the vQTL identified in this study by identifying vQTL intervals for putative candidate genes. We placed particular emphasis on genes that have annotations related to regulating mineral concentration in wheat and other plant species. For the vQTL on chromosome 2A, 38 candidate genes were identified in the 1.18 Mb interval that is physically located between 715,333,165 to 717,146,211 bp using IWGSC RefSeq v.1.0 (Supplementary File S2). For the vQTL on chromosome 2B, 108 candidate genes were predicted in the 9.32 Mb interval physically located from 691,780,716 to 701,097,263 bp based on IWGSC RefSeq v1.0. Based on the annotations for the identified candidate genes, many of the genes encoded homeobox-leucine zipper family protein, ABC transporter, MADS-box transcription factor, plant peroxidase, and glycosyltransferase, which have been associated with the genetic regulation of minerals in plants (Whitt et al., 2018). A shortlist of potential candidate genes is provided in Table 1, and the complete list can be found in Supplementary File S2. The results clearly showed that the two genomic regions associated with variance heterogeneity on chromosomes 2A and 2B harbor numerous putative candidate genes that potentially play significant roles in the genetic regulation of grain Cd concentration in wheat. However, we contend that further investigation of these regions using dense markers and increased sample size is necessary to fine-map the QTL and identify the causal genes underlying variation in these loci.

Discussion

In the present study, we explored the genetic variants affecting variance heterogeneity of Cd. Given the complexity of genetic regulation of Cd in wheat (Guttieri et al., 2015) and the influence of epistatic interactions, we anticipated that partial genetic regulation of Cd in wheat can be detected using methods that have been developed to identify vQTL. As reported by Rönnegård and Valdar (2011), a potential explanation for variance-controlling QTL is epistatic interactions that are unspecified in the model. Herein, we utilized two approaches, namely, DGLM and HGLM, to detect vQTL and mvQTL associated with grain Cd concentration in wheat.

The DGLM framework is a powerful approach for vGWAS analysis (Hulse and Cai, 2013). However, in DGLM, GLM is fitted by including only the fixed effects in the linear predictor of mean and dispersion. Therefore, by using the DGLM approach, population structure can only be accounted for by using the first few PCs obtained from the SNP matrix; however, this may not completely account for complex population structure and family relationships (Price et al., 2010). We hypothesized that the use of random effects to model the mean component can better account for population structure and reduce spurious associations. In this approach, a random additive genetic effect is introduced to the mean component of the model that accounts for population structure and cryptic relatedness between accessions. Therefore, we performed vGWAS analysis using HGLM. Interestingly, both DGLM and HGLM approaches were effective in identifying the genetic variants controlling variability of Cd, suggesting that the loci detected with the DGLM approach are likely to be true QTL rather than artifacts from population structure. The impact of population structure on the power of DGLM and HGLM remains to be explored; further examination is warranted.

In the literature, it has been argued that variance heterogeneity can also arise by a simple mean–variance relationship, which does not have biological significance (Young et al., 2018). To rule out the role of the mean-variance function in generating variance heterogeneity, we plotted the estimated effects of the top three significant associated markers at the alternate genotypes and observed that the means of all the markers were the same (Figure 3), indicating that the effect of SNP on variance heterogeneity was not due to the consequences of mean–variance function but likely due to the genetic effects (Yang et al., 2012).

In QTL studies, variance heterogeneity arises because of various underlying mechanisms, such as epistatic interactions (Struchalin et al., 2012; Shen et al., 2012; Nelson et al., 2013). Epistasis gives rise to variance heterogeneity when the different allele combinations at one locus change the effect of the other loci in the genome, as shown in one pair of interacting markers (Figure 4). Hence, identifying the loci affecting variance heterogeneity through vGWAS means that the loci are likely to be involved in epistatic interactions. To validate this assumption and investigate whether epistasis can explain the identified vQTL and mvQTL in this study, we analyzed all possible pairwise interactions between the associated markers. We detected significant epistatic interactions between the associated markers (Figure 2), which can explain the existence of variance heterogeneity in the genotypes. Additionally, identifying vQTL through vGWAS serves as an effective way to restrict the search space when detecting epistatic QTL. Thus, with the vGWAS approach, many of the requirements necessary for conventional epistasis mapping can be avoided (e.g., large sample size and extensive multiple testing corrections that reduce power). However, Forsberg and Carlborg (2017) empirically showed that the presence of variance heterogeneity does not always guarantee the presence of epistatic interactions that contribute to the total variation of the trait; therefore, the results should be interpreted carefully when multi-locus interactions are involved. Further, variance heterogeneity can also be observed in a population when two or more alleles having different effects on the phenotype are in high LD (Cao et al., 2014; Forsberg and Carlborg, 2017; Wang et al., 2014a). To rule out the possibility of LD as a source for variance heterogeneity in grain Cd in this population, we suggest the use of high-density markers and larger sample size to identify the actual functional alleles associated with Cd, their LD patterns, and their effects on the Cd phenotype (Struchalin et al., 2012; Forsberg and Carlborg, 2017).

We performed candidate gene analysis of the identified vQTL on chromosomes 2A and 2B to further explore the identified vQTL regions and elucidate the molecular basis underlying the Cd levels from these regions. The 2A and 2B regions were found to harbor numerous putative candidate genes encoding proteins with known functions (Table1 and Supplementary File S2). Some of the candidate genes included homeobox-leucine zipper family protein, ABC transporter, MADS-box transcription factor, plant peroxidase, and glycosyltransferase, all of which have been associated with genetic regulation of Cd in plants (Whitt et al., 2018). For instance, several metal transporters, including ABC transporters, play important roles in heavy metal uptake, transport, and distribution and play key roles in Cd tolerance (Wang et al., 2017; Zhu et al., 2018). ABC transporters have been associated with the regulation of Cd concentration in crops by inhibiting Cd uptake in roots, accumulation, transportation, and detoxification (Hu et al., 2019; Sheng et al., 2018; Zhang et al., 2018; Yao et al., 2018; Thakur et al., 2019; Wang et al., 2017). Similarly, homeodomain-leucine zipper family protein has been functionally associated with Cd tolerance by regulating the expression of metal transporters *OsHMA2* and *OsHMA3* in rice (Yu et al., 2019; Ding et al., 2018). These genes have been found to play important roles in loading Cd onto the xylem and root-to-shoot translocation of Cd in rice. In plants, response to heavy metals involves the accumulation of reactive oxygen species (ROS) that damage DNA and cellular machinery (Kumari et al., 2008; Rascio and Navari-Izzo, 2011). In *Arabidopsis*, the peroxidase genes *At2g35380*, *PER20*, and *At2g18150* have been found to be associated with Cd responses by affecting the lignin biosynthesis in root cells under high Cd stress (Mortel et al., 2008; Chen and Kao, 1995). The two genomic regions associated with variance heterogeneity harbor numerous putative candidate genes that are likely to play roles in regulating Cd concentrations in wheat. Further, the two genomic regions associated with variance heterogeneity presented sequence similarity and the 2A region falls within the 2B region (Supplementary File S2: Figure S4). This raises an important question whether the gene redundancy in polyploidy species has any role in generating the variance heterogeneity.

Conclusion

We showed the potential of vGWAS for dissecting the genetic architecture of complex traits and identifying novel genomic regions influencing variance heterogeneity in wheat. We provided evidence that many genes contribute to natural variation in grain Cd concentration through non-additive genetic effects. This is particularly evidenced by epistatic interactions between mvQTL on chromosome 5A and vQTL on chromosomes 2A and 2B.

Author's contributions

W.H. and G.M. conceived the study. W.H. performed the data analysis and drafted the manuscript. D.J. helped the data analysis. M.C., D.J., H.W., and G.M. revised the manuscript. G.M. supervised and directed the study. All authors read and approved the manuscript.

Acknowledgements

This work was supported by the National Science Foundation under Grant Number 1736192 to H.W. and G.M. Data analysis was performed using the Holland Computing Center computational resources at the University of Nebraska-Lincoln.

370 # Tables

Table 1: List of selected putative candidate genes based on function and literature search associated to variance heterogeneity in the genetic regulation of grain cadmium concentration in wheat.

Chr[1]	Gene ID	Putative Function	GO annotation[2]	Reference
2A	*TraesCS2A01G474000*	Homeobox-leucine zipper family protein	DNA binding	Zhu et al. (2018); Yu et al. (2019); Ding et al. (2018); Alomari et al. (2018)
2A/2B	*TraesCS2A01G474100/ TraesCS2B01G497600*	ABC transporter	Transporter activity; ATP binding; ATPase activity	Hu et al. (2019); Sheng et al. (2018); Zhang et al. (2018); Yao et al. (2018); Thakur et al. (2019)
2A	*TraesCS2A01G475000*	MADS-box transcription factor	Transcription factor activity, sequence-specific DNA binding; nucleus; regulation of transcription, DNA-templated	Yu et al. (2019); Zhao et al. (2019); Xu et al. (2018); Ding et al. (2018); Bhatta et al. (2018); Palmer et al. (2013)
2A/2B	*TraesCS2A01G476300/ TraesCS2B01G499900*	Peroxidase	Peroxidase activity; response to oxidative stress; oxidation-reduction process	Bhatta et al. (2018); Mortel et al. (2008)
2A/2B	*TraesCS2A01G474700/ TraesCS2B01G498300*	Glycosyltransferase	Metabolic process; transferase activity, transferring hexosyl groups	Xu et al. (2015); Peng et al. (2015)

371 Figures

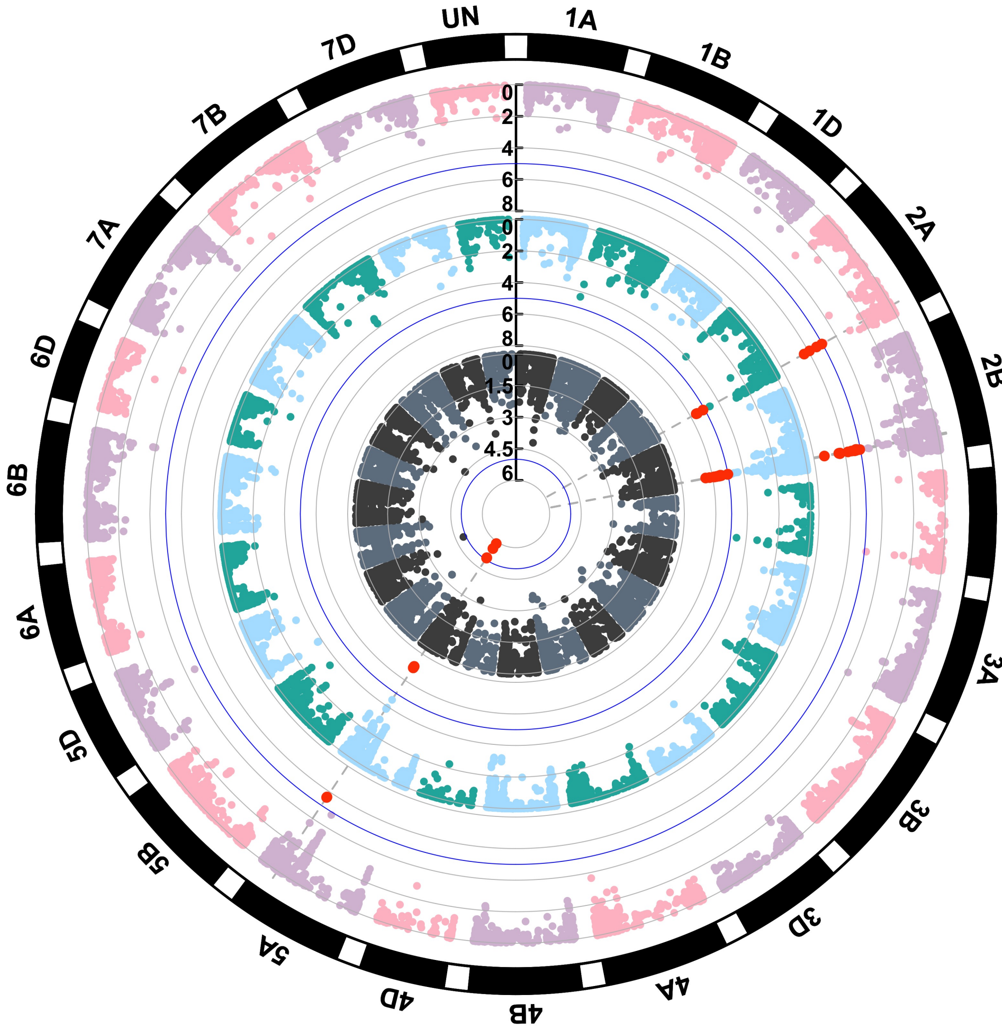

Figure 1: Circular Manhattan plot of standard genome-wide association studies (GWAS) based on mean differences (inner), and variance GWAS based on double generalized linear model (middle) and hierarchical generalized linear model (outer) for grain cadmium concentration in the hard-red winter wheat association panel. The red dots represent the significant markers associated with either mean or variance heterogeneity quantitative trait loci. The blue line in each circular plot shows the cutoff for the statistical significance ($P < 9.01\times10^{-6}$). The P-values in $-\log_{10}$ scale are given in black vertical line.

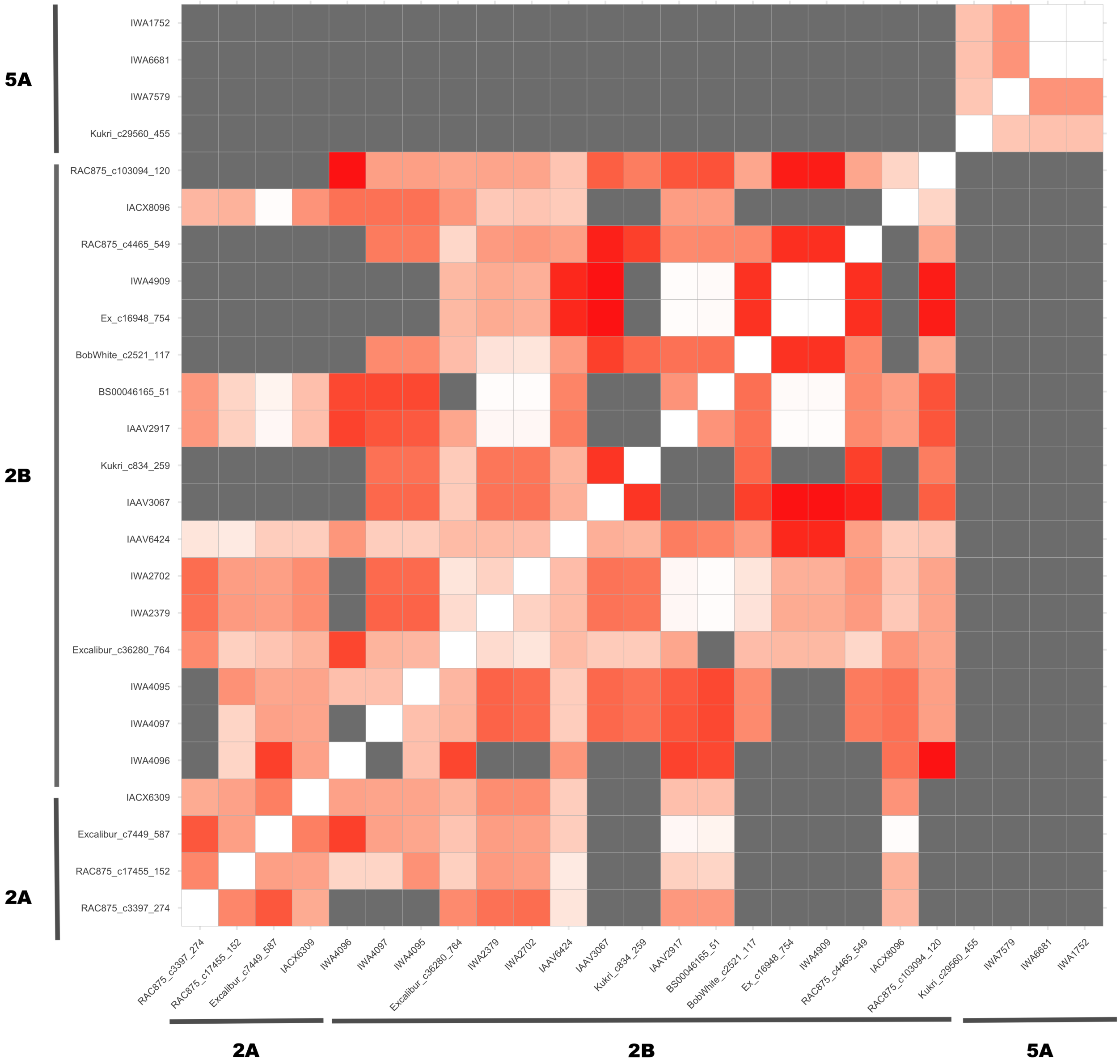

Figure 2: Heat map showing all possible pairwise epistatic interactions between the associated markers on chromosomes 2A, 2B, and 5A. The lower the P-value, the darker the shading. Interactions that are statistically significant ($P < 3.7 \times 10^{-5}$) are shown in gray color.

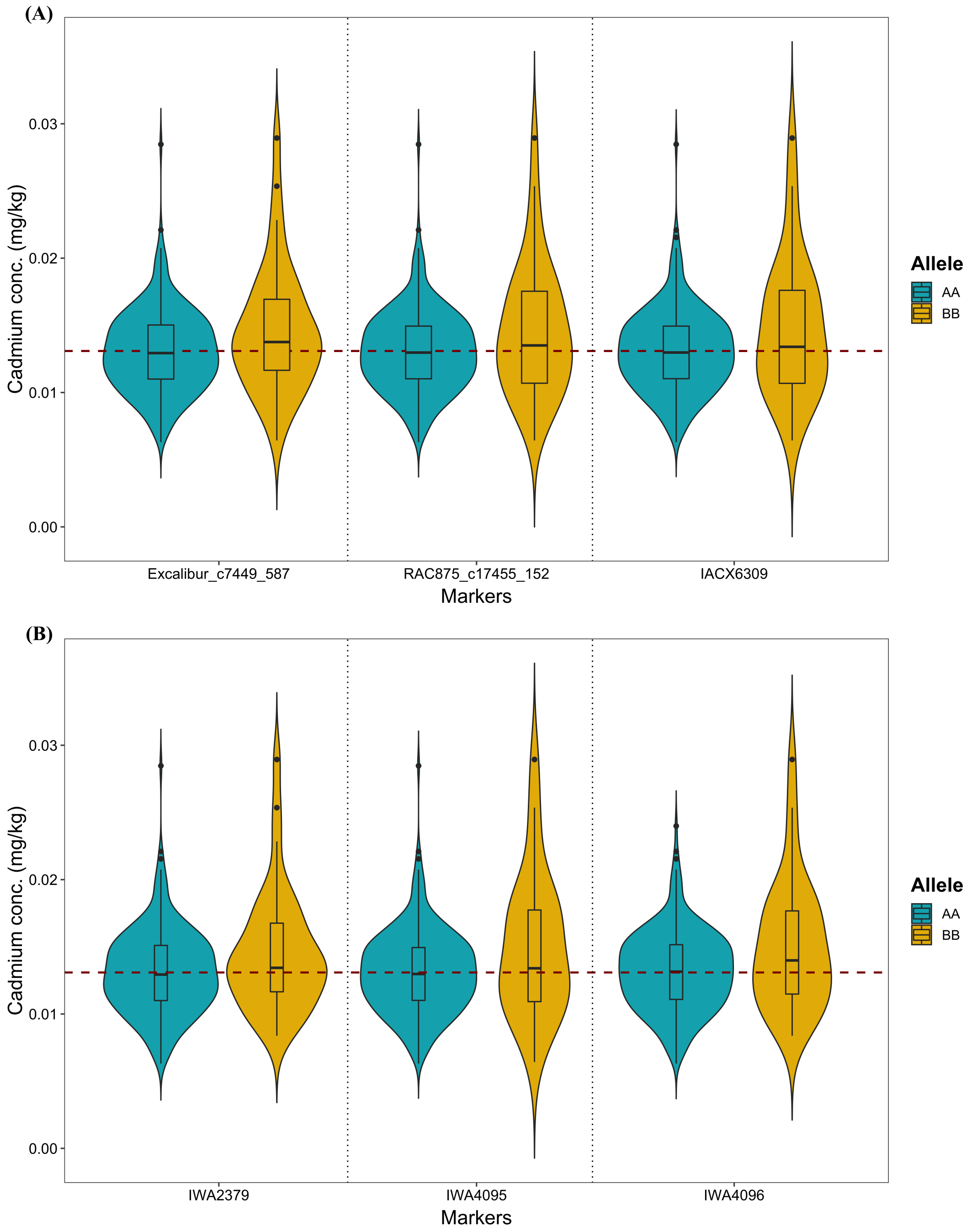

Figure 3: Violin plot showing the differences in the mean and variance of grain cadmium concentration with alternative marker allele groups coded as AA and BB for the top three significant markers associated with vQTL on (A) chromosome 2A and (B) chromosome 2B. The mean of marker genotypes AA and BB are connected by red dotted line.

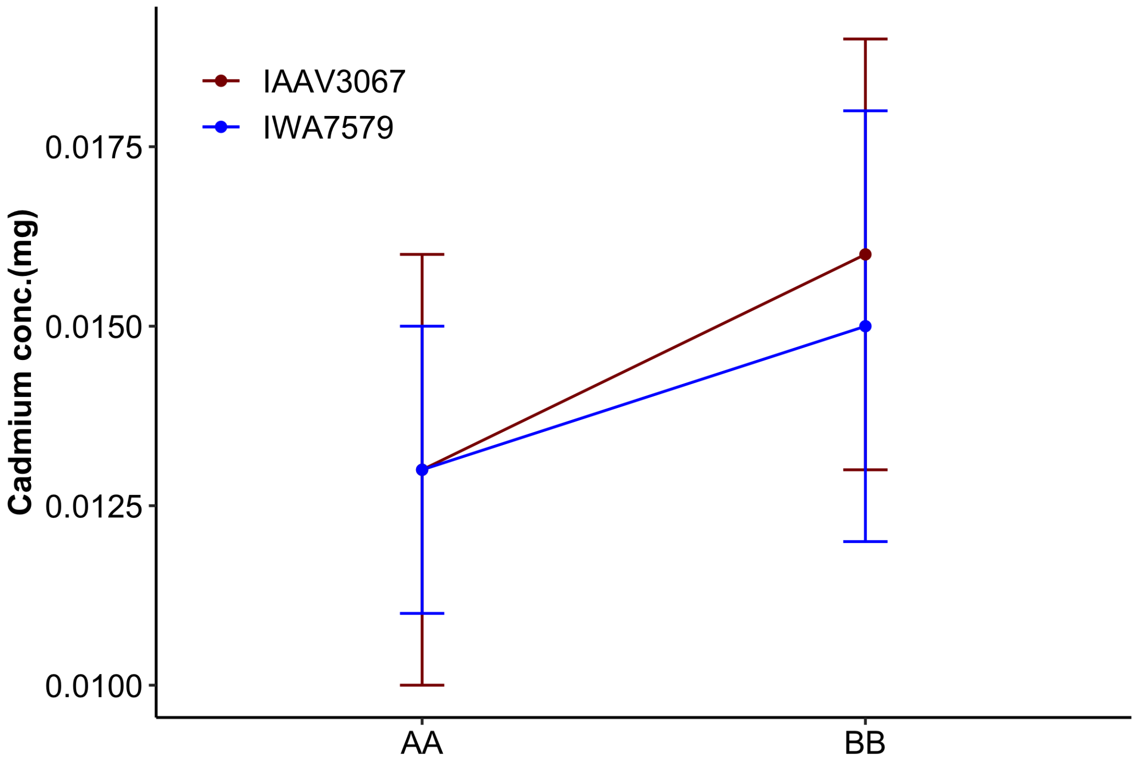

Figure 4: Epistatic interaction between single nucleotide polymorphisms on 5A (mvQTL) and 2B (vQTL) chromosomes. AA and BB represent the alternate genotypes at the particular SNP. Large difference in grain cadmium concentrations at BB genotype compared to no difference at AA genotype indicates the presence of interaction.

References

Alomari, D. Z., Eggert, K., Von Wirén, N., Alqudah, A. M., Polley, A., Plieske, J., Ganal, M. W., Pillen, K., and Röder, M. S. (2018). Identifying candidate genes for enhancing grain zn concentration in wheat. *Frontiers in Plant Science*, 9.

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 361(6403):eaar7191.

Ayroles, J. F., Buchanan, S. M., OLeary, C., Skutt-Kakaria, K., Grenier, J. K., Clark, A. G., Hartl, D. L., and Bivort, B. L. d. (2015). Behavioral idiosyncrasy reveals genetic control of phenotypic variability. *Proceedings of the National Academy of Sciences*, 112(21):6706–6711.

Bhatta, M., Baenziger, P. S., Waters, B. M., Poudel, R., Belamkar, V., Poland, J., and Morgounov, A. (2018). Genome-Wide Association Study Reveals Novel Genomic Regions Associated with 10 Grain Minerals in Synthetic Hexaploid Wheat. *International Journal of Molecular Sciences*, 19(10).

Bolser, D. M., Staines, D. M., Perry, E., and Kersey, P. J. (2017). Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. In *Plant Genomics Databases*, pages 1–31. Springer.

Brown, A. A., Buil, A., Viñuela, A., Lappalainen, T., Zheng, H.-F., Richards, J. B., Small, K. S., Spector, T. D., Dermitzakis, E. T., and Durbin, R. (2014). Genetic interactions affecting human gene expression identified by variance association mapping. *eLife*, 3:e01381.

Brown, M. B. and Forsythe, A. B. (1974). The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means. *Technometrics*, 16(1):129–132.

Cao, Y., Wei, P., Bailey, M., Kauwe, J. S., Maxwell, T. J., and Initiative, A. D. N. (2014). A versatile omnibus test for detecting mean and variance heterogeneity. *Genetic Epidemiology*, 38(1):51–59.

Chen, S. L. and Kao, C. H. (1995). Cd induced changes in proline level and peroxidase activity in roots of rice seedlings. *Plant Growth Regulation*, 17(1):67–71.

Corty, R. W., Kumar, V., Tarantino, L. M., Takahashi, J. S., and Valdar, W. (2018). Mean-Variance QTL Mapping Identifies Novel QTL for Circadian Activity and Exploratory Behavior in Mice. *G3: Genes, Genomes, Genetics*, page g3.200194.2018.

Corty, R. W. and Valdar, W. (2018). QTL mapping on a background of variance heterogeneity. *G3: Genes, Genomes, Genetics*, 8(12):3767–3782.

Ding, Y., Gong, S., Wang, Y., Wang, F., Bao, H., Sun, J., Cai, C., Yi, K., Chen, Z., and Zhu, C. (2018). Microrna166 modulates cadmium tolerance and accumulation in rice. *Plant Physiology*, 177(4):1691–1703.

Dumitrascu, B., Darnell, G., Ayroles, J., and Engelhardt, B. E. (2018). Statistical tests for detecting variance effects in quantitative trait studies. *Bioinformatics*.

Forsberg, S. K., Andreatta, M. E., Huang, X.-Y., Danku, J., Salt, D. E., and Carlborg, Ö. (2015). The multi-allelic genetic architecture of a variance-heterogeneity locus for molybdenum concentration in leaves acts as a source of unexplained additive genetic variance. *PLoS Genetics*, 11(11):e1005648.

Forsberg, S. K. G. and Carlborg, . (2017). On the relationship between epistasis and genetic variance heterogeneity. *Journal of Experimental Botany*, 68(20):5431–5438.

Guttieri, M. J., Baenziger, P. S., Frels, K., Carver, B., Arnall, B., Wang, S., Akhunov, E., and Waters, B. M. (2015). Prospects for selecting wheat with increased zinc and decreased cadmium concentration in grain. *Crop Science*, 55(4):1712–1728.

Hoffman, G. E. (2013). Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLoS ONE*, 8(10):e75707.

Hu, Y., Xu, L., Tian, S., Lu, L., and Lin, X. (2019). Site-specific regulation of transcriptional responses to cadmium stress in the hyperaccumulator, sedum alfredii: based on stem parenchymal and vascular cells. *Plant Molecular Biology*, pages 1–16.

Hulse, A. M. and Cai, J. J. (2013). Genetic variants contribute to gene expression variability in humans. *Genetics*, 193(1):95–108.

Kumari, M., Taylor, G. J., and Deyholos, M. K. (2008). Transcriptomic responses to aluminum stress in roots of arabidopsis thaliana. *Molecular Genetics and Genomics*, 279(4):339.

Kusmec, A., Srinivasan, S., Nettleton, D., and Schnable, P. S. (2017). Distinct genetic architectures for phenotype means and plasticities in zea mays. *Nature Plants*, 3(9):715.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4):619–656.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized linear models with random effects: unified analysis via H-likelihood*. Chapman and Hall/CRC.

Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221.

Mackay, T. F. and Lyman, R. F. (2005). Drosophila bristles and the nature of quantitative genetic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1513–1527.

Menke, A., Muntner, P., Silbergeld, E. K., Platz, E. A., and Guallar, E. (2008). Cadmium levels in urine and mortality among US adults. *Environmental Health Perspectives*, 117(2):190–196.

Mortel, J. E. V. D., Schat, H., Moerland, P. D., Themaat, E. V. L. V., Ent, S. V. D., Blankestijn, H., Ghandilyan, A., Tsiatsiani, S., and Aarts, M. G. M. (2008). Expression differences for genes involved in lignin, glutathione and sulphate metabolism in response to cadmium in Arabidopsis thaliana and the related Zn/Cd-hyperaccumulator Thlaspi caerulescens. *Plant, Cell & Environment*, 31(3):301–324.

Nachman, M. W., Hoekstra, H. E., and D'Agostino, S. L. (2003). The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences*, 100(9):5268–5273.

Nelson, R. M., Pettersson, M. E., Li, X., and Carlborg, . (2013). Variance Heterogeneity in Saccharomyces cerevisiae Expression Data: Trans-Regulation and Epistasis. *PLoS ONE*, 8(11):e79507.

Palmer, C. M., Hindt, M. N., Schmidt, H., Clemens, S., and Guerinot, M. L. (2013). Myb10 and myb72 are required for growth under iron-limiting conditions. *PLoS Genetics*, 9(11):e1003953.

Par, G., Cook, N. R., Ridker, P. M., and Chasman, D. I. (2010). On the Use of Variance per Genotype as a Tool to Identify Quantitative Trait Interaction Effects: A Report from the Women's Genome Health Study. *PLoS Genetics*, 6(6):e1000981.

Patterson, N., Price, A. L., and Reich, D. (2006). Population Structure and Eigenanalysis. *PLOS Genetics*, 2(12):e190.

Peng, H., He, X., Gao, J., Ma, H., Zhang, Z., Shen, Y., Pan, G., and Lin, H. (2015). Transcriptomic changes during maize roots development responsive to cadmium (cd) pollution using comparative rnaseq-based approach. *Biochemical and Biophysical Research Communications*, 464(4):1040–1047.

Perry, G. M. L., Nehrke, K. W., Bushinsky, D. A., Reid, R., Lewandowski, K. L., Hueber, P., and Scheinman, S. J. (2012). Sex Modifies Genetic Effects on Residual Variance in Urinary Calcium Excretion in Rat (*Rattus norvegicus*). *Genetics*, 191(3):1003–1013.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459.

Queitsch, C., Sangster, T. A., and Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. *Nature*, 417(6889):618–624.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rascio, N. and Navari-Izzo, F. (2011). Heavy metal hyperaccumulating plants: how and why do they do it? and what makes them so interesting? *Plant Science*, 180(2):169–181.

Rönnegård, L. and Carlborg, Ö. (2007). Separation of base allele and sampling term effects gives new insights in variance component qtl analysis. *BMC Genetics*, 8(1):1.

Rönnegård, L., Shen, X., and Alam, M. (2010). hglm: A package for fitting hierarchical generalized linear models. *The R Journal*, 2(2):20–28.

Rönnegård, L. and Valdar, W. (2011). Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics*, 188(2):435–447.

Rönnegård, L. and Valdar, W. (2012). Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC Genetics*, 13(1):63.

Salom, P. A., Bomblies, K., Laitinen, R. A. E., Yant, L., Mott, R., and Weigel, D. (2011). Genetic Architecture of Flowering-Time Variation in Arabidopsis thaliana. *Genetics*, 188(2):421–433.

Sell-Kubiak, E., Duijvesteijn, N., Lopes, M., Janss, L., Knol, E., Bijma, P., and Mulder, H. (2015). Genome-wide association study reveals novel loci for litter size and its variability in a large white pig population. *BMC Genomics*, 16(1):1049.

Shen, X., Pettersson, M., Rönnegård, L., and Carlborg, Ö. (2012). Inheritance beyond plain heritability: variance-controlling genes in arabidopsis thaliana. *PLoS Genetics*, 8(8):e1002839.

Sheng, Y., Yan, X., Huang, Y., Han, Y., Zhang, C., Ren, Y., Fan, T., Xiao, F., Liu, Y., and Cao, S. (2018). The wrky transcription factor, wrky13, activates pdr8 expression to positively regulate cadmium tolerance in arabidopsis. *Plant, Cell & Environment*.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 47–60.

Struchalin, M. V., Amin, N., Eilers, P. H., Duijn, C. M. v., and Aulchenko, Y. S. (2012). An R package ”VariABEL” for genome-wide searching of potentially interacting loci by testing genotypic variance heterogeneity. *BMC Genetics*, 13(1):4.

Sul, J. H., Martin, L. S., and Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genetics*, 14(12):e1007309.

Tan, Q., Hjelmborg, J. V., Thomassen, M., Jensen, A. K., Christiansen, L., Christensen, K., Zhao, J. H., and Kruse, T. A. (2014). Hierarchical linear modeling of longitudinal pedigree data for genetic association analysis. In *BMC Proceedings*, volume 8, page S82. BioMed Central.

Thakur, S., Choudhary, S., and Bhardwaj, P. (2019). Comparative transcriptome profiling under cadmium stress reveals the uptake and tolerance mechanism in brassica juncea. *Journal of Plant Growth Regulation*, pages 1–12.

Topless, R. K., Flynn, T. J., Cadzow, M., Stamp, L. K., Dalbeth, N., Black, M. A., and Merriman, T. R. (2015). Association of slc2a9 genotype with phenotypic variability of serum urate in pre-menopausal women. *Frontiers in Genetics*, 6:313.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423.

Wang, G., Yang, E., Brinkmeyer-Langford, C. L., and Cai, J. J. (2014a). Additive, epistatic, and environmental effects through the lens of expression variability QTL in a twin cohort. *Genetics*, 196(2):413–425.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S. G., Cattivelli, L., et al. (2014b). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnology Journal*, 12(6):787–796.

Wang, Y., Wang, X., Wang, C., Peng, F., Wang, R., Xiao, X., Zeng, J., Kang, H., Fan, X., Sha, L., et al. (2017). Transcriptomic profiles reveal the interactions of cd/zn in dwarf polish wheat (triticum polonicum l.) roots. *Frontiers in Physiology*, 8:168.

Wei, W.-H., Bowes, J., Plant, D., Viatte, S., Yarwood, A., Massey, J., Worthington, J., and Eyre, S. (2016). Major histocompatibility complex harbors widespread genotypic variability of non-additive risk of rheumatoid arthritis including epistasis. *Scientific Reports*, 6:25014.

Wei, W.-H., Massey, J., Worthington, J., Barton, A., and Warren, R. B. (2018). Genotypic variability-based genome-wide association study identifies non-additive loci hla-c and il12b for psoriasis. *Journal of Human Genetics*, 63(3):289.

Wei, W.-H., Viatte, S., Merriman, T. R., Barton, A., and Worthington, J. (2017). Genotypic variability based association identifies novel non-additive loci DHCR7 and IRF4 in seronegative rheumatoid arthritis. *Scientific Reports*, 7(1):5261.

Whitt, L., Ricachenevsky, F., Ziegler, G., Clemens, S., Walker, E., Maathuis, F. J. M., Kear, P., and Baxter, I. (2018). A curated list of genes that control elemental accumulation in plants. *bioRxiv*, page 456384.

Xiao, Y., Liu, H., Wu, L., Warburton, M., and Yan, J. (2017). Genome-wide Association Studies in Maize: Praise and Stargaze. *Molecular Plant*, 10(3):359–374.

Xu, L., Wang, Y., Liu, W., Wang, J., Zhu, X., Zhang, K., Yu, R., Wang, R., Xie, Y., Zhang, W., et al. (2015). De novo sequencing of root transcriptome reveals complex cadmium-responsive regulatory networks in radish (raphanus sativus l.). *Plant Science*, 236:313–323.

Xu, N., Chu, Y., Chen, H., Li, X., Wu, Q., Jin, L., Wang, G., and Huang, J. (2018). Rice transcription factor osmads25 modulates root growth and confers salinity tolerance via the aba–mediated regulatory pathway and ros scavenging. *PLoS Genetics*, 14(10):e1007662.

Yang, J., Loos, R. J., Powell, J. E., Medland, S. E., Speliotes, E. K., Chasman, D. I., Rose, L. M., Thorleifsson, G., Steinthorsdottir, V., Mägi, R., et al. (2012). FTO genotype is associated with phenotypic variability of body mass index. *Nature*, 490(7419):267.

Yao, X., Cai, Y., Yu, D., and Liang, G. (2018). bhlh104 confers tolerance to cadmium stress in arabidopsis thaliana. *Journal of Integrative Plant Biology*.

Young, A. I., Wauthier, F. L., and Donnelly, P. (2018). Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nature Genetics*, 50(11):1608.

Yu, J., Wu, L., Fu, L., Shen, Q., Kuang, L., Wu, D., and Zhang, G. (2019). Genotypic difference of cadmium tolerance and the associated micrornas in wild and cultivated barley. *Plant Growth Regulation*, pages 1–13.

Zhang, X. D., Zhao, K. X., and Yang, Z. M. (2018). Identification of genomic atp binding cassette (abc) transporter genes and cd-responsive abcs in brassica napus. *Gene*, 664:139–151.

Zhao, J., Yuan, S., Zhou, M., Yuan, N., Li, Z., Hu, Q., Bethea, F. G., Liu, H., Li, S., and Luo, H. (2019). Transgenic creeping bentgrass overexpressing Osa-miR393a exhibits altered plant development and improved multiple stress tolerance. *Plant Biotechnology Journal*, 17(1):233–251.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of snp data. *Bioinformatics*, 28(24):3326–3328.

Zhu, H., Ai, H., Cao, L., Sui, R., Ye, H., Du, D., Sun, J., Yao, J., Chen, K., and Chen, L. (2018). Transcriptome analysis providing novel insights for cd-resistant tall fescue responses to cd stress. *Ecotoxicology and Environmental Safety*, 160:349–356.