Constraints to Adaptation in Maize: Environmental Trade-offs and Deleterious Alleles

By

Asher I. Hudson

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Population Biology

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Jeffrey Ross-Ibarra, Chair

Daniel Runcie

Johanna Schmitt

Committee in Charge

2022

Contents

Abstract

Constraints to Adaptation in Maize: Environmental Trade-offs and Deleterious Alleles

Selection can result in populations becoming more adapted, yet even in the presence of selection other effects can constrain adaptation. Two of the effects that can cause this discrepancy are genotype by environment interactions (GxE) and the accumulation of deleterious alleles. GxE is observed when an allele has different effects depending on the environment and can result in the maintenance of genetic variation, particularly when no genotype is best adapted to all environments. Deleterious alleles are generally considered as those which are deleterious across the conditions members of a species might encounter. Although they experience negative selection, they can still contribute a substantial proportion of genetic variance.

In this dissertation, I analyzed GxE and deleterious alleles in maize. First, I investigated GxE in a maize mapping population. We find that GxE contributes a substantial amount to the phenotypic variance for many traits. While we identify loci contributing to GxE, overall most of the GxE variance may be due to unidentified polygenic effects. Estimating the genetic covariances between traits in each environment reveals large differences in the genetic variance-covariance matrix between environments and in particular shows that differences in selection on flowering time may be contributing to the observed GxE for yield. Second, we analyzed the distribution of structural variants in maize inbred lines along the genome. We find that structural variants are more depleted in constrained regions of the genome than single nucleotide polymorphisms, possibly indicating that structural variants are more likely to be deleterious. Finally, we apply a machine learning method to identify constrained regions of the maize genome based on population genetic data. These predictions allow us to assess more recent evolutionary constraint in maize and find regions where mutations are more likely to be deleterious.

Acknowledgments

I would like to thank Jeffrey Ross-Ibarra and the Ross-Ibarra lab for intellectual community, supporting me in both research and outreach, and convincing me that crops are an interesting thing for an evolutionary biologist to study. I have received invaluable advice from my dissertation committee, Daniel Runcie and Johanna Schmitt, as well as the remaining members of my qualifying examination committee, Amélie Gaudin, Sharon Strauss, and Michael Turelli. The research in this dissertation was all carried out in collaboration with other scientists: thank you to Sebastien Praud, Marie-Helene Tixier, and Pierre Dubreuil at Limagrain, Matthew Hufford and Arun Seetharam at Iowa State University, and my labmate Sarah Odell. The Population Biology Graduate Group has been an amazing community to be a part of. While every member of the graduate group has contributed to this community, in particular I would like to thank my cohort, Katherine Corn, Elise Elwood, Fernanda Guizar, and Sivan Yair, as well as Michelle Stitzer and Anita To. Finally, thank you to my parents, Amy Cherry and Avery Hudson, my partner, Maxfield Ellinger, and my cat Lionelle, who encourages work-life balance by regularly sitting directly on my laptop while I am trying to type.

Chapter 1

Analysis of genotype by environment interactions in a maize mapping population

Asher I. Hudson[1,2], Sarah G. Odell[1,3], Pierre Dubreuil[4], Marie-Helene Tixier[4], Sebastien Praud[4], Daniel E. Runcie[3], Jeffrey Ross-Ibarra[1,2,5]

[1] Department of Evolution and Ecology, University of California, Davis, CA, USA

[2] Center for Population Biology, University of California, Davis, CA USA

[3] Department of Plant Sciences, University of California, Davis, CA, USA

[4] Center of Research of Chappes, Limagrain, Chappes, France

[5] Genome Center, University of California, Davis, CA, USA

1.1 Introduction

Both the effect of a given genotype on a trait, and the impact of that effect on fitness, often vary across environments. Such genotype by environment interactions (GxE) are widespread, and have been commonly observed in plants (Bradshaw, 1965; Des Marais *et al.*, 2013). GxE interactions are of interest for multiple reasons: they provide insight into the physiological processes and genetic architecture underlying individual traits, are likely crucial for local adaptation of populations to different environments, but may also limit the response to selection (Allard and Bradshaw, 1964; Kawecki and Ebert, 2004).

While alleles affecting a trait will demonstrate GxE for fitness across environments when there is selection for different trait optima, it is also often observed that the effect of individual alleles on traits will vary as well. This indicates that these alleles affect plasticity and they may be present in a population due to selection for or against plasticity (Josephs, 2018). Alternatively, they may be deleterious but rarely exposed to environments in which they are selected against, or unassociated with fitness and selectively neutral (Des Marais *et al.*, 2013; Paaby and Rockman, 2014).

One avenue to study GxE is to search for individual loci with changing effects on traits or fitness across environments. Multiple studies have identified loci that contribute to GxE (several of which are reviewed in Josephs (2018)). Loci which contribute to GxE include the Eda locus in threespine stickleback fish, which is associated with adaptation to the freshwater environment, and Sub1A in rice, which is associated with tolerance to submergence (Barrett *et al.*, 2008; Xu *et al.*, 2006). Genome-wide association studies (GWAS) have also been used to identify alleles significantly associated with GxE, including shade response and drought response in *Arabidopsis thaliana* (Filiault and Maloof, 2012; El-Soda *et al.*, 2015).

Individual traits do not exist in a vacuum, however, and alleles that affect one trait often have pleiotropic effects on others. Indeed, the outcome of selection on a trait depends crucially on the genetic variance-covariance matrix (G-matrix), which describes how the genetic value at one trait covaries with genetic values at other traits (Lande, 1979). Genetic covariation between traits can have profound impacts on the genetic response to selection, either hindering or facilitating trait response. For example, if fitness positively co-varies with two different traits, but those traits negatively co-vary with each other, this can lead to a trade-off.

But the G-matrix itself is not constant, as GxE at underlying loci may impact trait variation and covariation among traits (Wood and Brodie, 2015). If in a different environment the covariance of a trait with fitness or other traits is weakened or changes sign, it may indicate that the selection or trade-off does not exist in the new environment (Sgrò and Hoffmann, 2004). As GxE contributes to the G-matrix within each environment, understanding the G-matrix in multiple environments may illuminate the causes of GxE. If the genetic covariance between two traits changes between environments and GxE is observed, then a change in the pleiotropy of the underlying loci may be responsible for both the changes in the genetic covariance and GxE.

Maize is a crop species adapted to a wide diversity of environments, from temperate to tropical and from low to high altitude (Hake and Ross-Ibarra, 2015). GxE has been shown to be an important contributor to many traits in maize, including grain yield (Gage *et al.*, 2017; Gates *et al.*, 2019; Rogers *et al.*, 2021). Nonetheless, identification of GxE in maize, as in many species, is complicated by issues of population structure and the low minor allele frequency of most polymorphisms (Korte and Farlow, 2013). To circumvent these issues, we investigated the genetic basis of GxE in maize in a multiparent advanced generation intercross (MAGIC) population of 16 diverse temperate maize lines (Odell *et al.*, 2022). We grew the MAGIC hybrids across five contrasting temperate environments with diverse management practices in order to capture a broad range of GxE relevant to the conditions the parental lines would be grown in.

We find that GxE contributes as much as genotypic main effects to variance for some traits. While GxE interactions are significant, genome-wide association only finds a small number of markers significantly associated with GxE interactions, perhaps reflecting the highly polygenic nature of most traits. Nonetheless, estimation of the G-matrix in each environment reveals that changes in genetic covariance are common and may be contributing to observed GxE. For example, we find that while only a small proportion of variance in flowering time depends on GxE, the genetic covariance between flowering time and grain yield is strongly affected by the environment.

1.2 Materials and methods

1.2.1 Plant materials

We developed a maize multi parent advanced generation intercross (MAGIC) population by repeatedly crossing the offspring of sixteen maize inbred lines to generate recombinant individuals (Odell *et al.*, 2022). Inbred lines were selected to maximize genetic diversity and include dent, flint, and European flint lines. After eight generations of intercrossing, we generated a population of 344 doubled haploids (DH) lines. DH lines were crossed to MBS847, a dent line chosen to be the tester, to make F1 plants.

1.2.2 Phenotype Data

The MAGIC F1 plants were phenotyped in four different field locations in four different years, resulting in five distinct environment-years (Supplemental fig. 1.4, supplemental table 1.1). The environment-years included Blois, France in 2014 and 2017, Nerac, France in 2016, St. Paul, France in 2017, and Graneros, Chile in 2015. We used an alpha design with two plots of around 80 plants grown for each genotype in each environment-year. Planting density ranged between 85,000 to 95,000 seeds per hectare. Seeds were planted with an automatic seed drill. The row width was 0.8 meters with two rows per plot. The fields in environment-years Blois 2014, Blois 2017, and Graneros 2015 all received consistent irrigation. The field in Nerac 2016 was not actively irrigated from vegetative phase through flowering, causing drought stress through most of the life cycle. The field in St. Paul 2017 was not irrigated during vegetative phase but was irrigated during flowering to allow plants to recover from the earlier drought stress. The applied drought stress was mild and intended to be representative of realistic field conditions.

We measured the following traits: male flowering date, female flowering date, anthesis-silking interval (ASI), plant height, percent harvest grain moisture (HGM), grain yield, and thousand kernel weight (TKW), where values were averaged over plots. Both flowering time phenotypes were measured as the sum of degree days since sowing with a base temperature of 6°C (48°F). Male flowering date was considered as the growing degree days (GDD) until 50% of plants in a plot were shedding pollen on approximately one quarter of the central tassel spike. Female flowering date was considered as the GDD until 50% of plants in a plot were flowering with 2 cm of silk outside of husk leaves. Plant height was measured as the distance from the base of the plant to the top of the tassel. Grain was collected using a combine harvest. Grain yield and TKW were both adjusted to 15% humidity. TKW was estimated from a 100 kernel sample. Data was also collected from an additional environment, Szeged, Hungary in 2017. We did not use this data in the analyses presented here as flowering date was not collected on the same schedule as in the other environments and this caused issues with the GxE analyses. Data from Szeged is available in the data repository associated with this paper. Between 292 and 309 of the MAGIC F1 lines were grown in each environment. There were a total of 325 lines that had both genotype data and phenotype data from at least one environment.

1.2.3 Genotyping

We genotyped each of the DH lines using the Affymetrix® Axiom® Maize Genotyping Array, which successfully genotyped 551,460 SNPs. The probability of each founder contributing to each segment in the genome was imputed from the genotyped SNPs (Odell *et al.*, 2022).

1.2.4 Estimating kinship

Kinship matrices for the DH lines were estimated from the genotyped SNPs using the Van-Raden method as implemented in the R package *sommer* (Covarrubias-Pazaran, 2016; Van-Raden, 2008; R Core Team, 2020). SNPs were first filtered for linkage disequilibrium using Plink with a window size of 50 kb, a step size of 5, and an r^2 threshold of 0.2 (Purcell *et al.*, 2007). In order to perform genome-wide association analyses, we used the leave one chromosome out method (Lippert *et al.*, 2011).

1.2.5 Genotype x environment interactions

Variance components for each trait were estimated using the R package *sommer*. We used the formula:

$$\mathbf{y} = \mathbf{Z_G u_G} + \mathbf{Z_E u_E} + \mathbf{Z_{E:G} u_{E:G}} + f_E(\mathbf{x}, \mathbf{y}) + e$$

Where $\mathbf{y}$ is a vector of n observations from individual plots of a single trait including both plots of all lines in all environments, $\mathbf{Z_G}$ is a $n \times r$ design matrix for the genotypic main effects of the r lines, $\mathbf{Z_E}$ is a $n \times 5$ design matrix for the environmental main effect, $\mathbf{Z_{E:G}}$ is a $n \times 5r$ design matrix for genotype specific effects in each environment, $\mathbf{u_G}$ is a length r vector of random genotypic effects, $\mathbf{u_E}$ is a length 5 vector of environmental random effects, $\mathbf{u_{E:G}}$ is a length $5r$ vector of random GxE effects with same variance and covariance among environments, $f_E(\mathbf{x}, \mathbf{y})$ is a two dimensional spline for the effect of the x/y position in the field nested within environment modeled as a single random effect fit from an incidence matrix containing the tensor products of the x and y coordinates in the field, and $\mathbf{e}$ is the error. *sommer* models 2D splines based on modified code from *SpATS* (Rodríguez-Álvarez *et al.*, 2017).

1.2.6 GWAS

Genome-wide association analyses for loci contributing to GxE interactions were performed with the R package *GridLMM* (Runcie and Crawford, 2019). Imputed founder probabilities at each locus were used as markers, meaning that at each marker we asked if the identity of the founder which contributed that genomic region at a given locus was a significant predictor of differences in plasticity among the hybrids. We set GridLMM to obtain maximum likelihood estimates of the effect of each marker.

GxE models can be parameterized in multiple ways which could potentially capture different aspects of GxE. We chose to model GxE in three different ways in our GWAS analyses, which we describe below.

i) Main effect across environments and deviation effect within environments

We tested whether a locus had a different effect on a trait in two environments: Blois 2017 and Nerac 2016. We chose these two environments because they were respectively the highest and lowest yielding environments. The model for this GWA was: y=$\mu + \mathbf{w}\alpha + \mathbf{X}_m\beta_m + \mathbf{X}_{E:m}\beta_{E:m} + \mathbf{Z}_{G1}\mathbf{u}_{G1} + \mathbf{Z}_{E:G1}\mathbf{u}_{E:G1} + \mathbf{Z}_{G2}\mathbf{u}_{G2} + \mathbf{e}$

Where $\mathbf{y}$ is a vector of n observations from individual plots of a single trait including both plots of all lines in both environments, μ is a constant length n vector of the average trait value across the two environments, $\mathbf{w}$ is a length n design matrix of environmental effects taking values of -1 and +1 according to the environment (1 for Blois 2017 and -1 for Nerac 2016), α is a scalar representing $\frac{1}{2}$ the deviation of trait means between the two environments, $\mathbf{X}_m$ is a $n \times 16$ matrix, where the kth column is the probability that each of the n individuals inherited from the kth founder at marker m, $\mathbf{X_{E:m}}$ is an n x 16 matrix formed by multiplying $\mathbf{w}$ with each column of $\mathbf{X}_m$, β_m is a vector of main effects of the founder alleles averaged over the two environments, $\beta_{E:m}$ is a vector of differences between the founder allele effects between the two environments, $\mathbf{Z}_{G1}$ is a $n \times r$ design matrix of additive genotypic effects, $\mathbf{Z}_{E:G1}$ is a $n \times r$ design matrix of genotype deviations formed by multiplying each column of $\mathbf{Z}_{G1}$ by $\mathbf{w}$, $\mathbf{Z}_{G2}$ is a $n \times r$ design matrix of non-additive genotypic effects, $\mathbf{u}_{G1}$ is a vector of additive genotypic effects averaged over the two environments, $\mathbf{u}_{E:G1}$ is a vector of additive genotypic deviations between the two environments, $\mathbf{u}_{G2}$ is a vector of non-additive genotypic effects averaged across the two environments, and $\mathbf{e}$ is a vector of error terms. $\mathbf{u}_{G1}$ and $\mathbf{u}_{E:G1}$ both have covariance proportional to $\mathbf{K}$, where $\mathbf{K}$ is the additive genetic relatedness matrix, and $\mathbf{u}_{G2}$ and $\mathbf{e}$ both have covariance proportional to the identity matrix. The statistical test to identify markers influencing GxE was against H0: $\beta_{E:m} = \mathbf{0}$.

ii) Plasticity

We tested whether a locus had an effect on the slope of the observations of a genotype across the mean phenotypic value of all genotypes in an environment. This model has the benefit of including the maximum amount of data. Compared to the main effect and deviation model (i), this model might be more likely to pick up GxE effects that have smaller effects within those two environments but a larger effect on the overall slope across environments. The model is the same as in i) except for the following: we now include all 5 environments, $\mathbf{w}$ is a length n vector with each element taking the mean value of the phenotype within the environment of the observation, and μ is a length n vector of the mean value of the phenotype within the environment of the observation.

iii) Finlay-Wilkinson GWAS

Finally, we tested whether a locus had an effect on the slope of the observations of a genotype across the mean grain yield of all genotypes in an environment. Mean grain yield here serves as a proxy for stress or environment quality and as such this GWA is testing whether a locus affects the response to stress. This is known as a Finlay-Wilkinson analysis (Finlay and Wilkinson, 1963). For this analysis, a quantile plot of p-values indicated that the test was poorly calibrated. Instead of asking whether allowing a marker to have a slope across environments improved prediction of a trait in each environment as in (ii), we thus asked whether the marker significantly predicted the slope of each genotype.

$$\mathbf{s} = \mathbf{X}_m\beta_s + \mathbf{Z}_{G1}\mathbf{u}_s + \mathbf{e}$$

Where $\mathbf{s}$ is a length r vector of slopes for each genotype of trait values on mean grain yield in each environment, β_s is a vector of marker effects, and $\mathbf{u}_s$ is a vector of genotypic effects with covariance proportional to $\mathbf{K}$. Other model terms are as in (i).

To determine significance thresholds for the first two models, we permuted phenotypic values among lines within each environment and ran the GWA 100 times. For the third model, we permuted the slopes among the genotypes and ran the GWA 100 times.

1.2.7 The G-matrix across environments

We estimated the G-matrix in each environment using the R package *brms* (Bürkner, 2017). *brms* implements Bayesian multilevel models using Markov chain Monte Carlo (MCMC) algorithms. This is important as the samples from the MCMC chains allow us to estimate uncertainty and significance in our downstream analyses. We used the model:

$$\mathbf{Y} = \mathbf{ZU} + f(\mathbf{x}, \mathbf{y}) + \mathbf{E}$$

Where $\mathbf{Y} = [\mathbf{y_1} ... \mathbf{y_5}]$ and $\mathbf{y_i}$ is a vector of n observations for the ith trait, $\mathbf{Z}$ is a $n \times r$ design matrix of genotypes, $\mathbf{U}$ and $\mathbf{E}$ are random effects drawn from multivariate normal distributions: $vec(\mathbf{U}) \sim N(vec(0), \mathbf{G} \otimes \mathbf{I_r})$, $vec(\mathbf{E}) \sim N(vec(0), \mathbf{R} \otimes \mathbf{I_n})$, $\mathbf{I_r}$ is the $r \times r$ identity matrix where r is the number of lines grown in an environment, $\mathbf{I_n}$ is the n x n identity matrix where n is the number of observations, and $\mathbf{G}$ and $\mathbf{R}$ are 5×5 genetic variance-covariance and residual variance-covariance matrices estimated from the data. G and R are parameterized as the products of standard deviations and correlation matrices with a half Student-T distribution and LKJ-correlation prior. $f(\mathbf{x}, \mathbf{y})$ is a two dimensional spline for the effect of the x/y position in the field. The standard deviations of the two splines have half Student-T distributions as priors.

All traits were scaled by the mean value across all environments and centered before analysis in order to make them unitless and improve model convergence. We performed this same analysis with non-scaled traits so that our results can be compared with those of previous studies with non-scaled phenotypic data. The G-matrices we estimated were broad sense G-matrices as they included both additive and non-additive sources of genetic variance. We ran four chains with 1,500 iterations of burn-in followed by 3,500 iterations. We chose these numbers as the *brms* documentation states that most models will converge with only a few thousand iterations. We assessed convergence by checking that all statistics output by *brms* — such as $\hat{R}$, defined as the potential scale reduction factor on split chains, and the number of divergent transitions, which occur when the simulated trajectory along the posterior differs from the true trajectory — were within recommended ranges and by visually inspecting the trace and autocorrelation of model parameters. For genotypic standard deviations and correlations, the bulk effective sample size of parameters ranged from 1,506 to 6,449. To determine whether the correlation between two traits differed between environments, we found the difference between the MCMC samples for the two environments and determined whether the interval spanned by the 2.5% and 97.5% quantiles of the differences overlapped zero. In particular, if the correlation between two traits was positive in one environment and negative in another, and if one or both of those traits correlate with yield, this would be evidence for a possible trade-off between fitness in different environments.

To quantitatively assess differences among the G-matrices estimated in the five environments, we performed eigenanalysis of a covariance tensor as described in Aguirre et al. (Aguirre *et al.*, 2014). The tensor approach is a geometric approach founded on the diagonalization of symmetric matrices, and is mainly used to calculate a set of orthogonal axes known as eigentensors that describe coordinated changes in the elements of the original matrices being compared. Eigentensors describe which elements of a set of matrices most contribute to variation among those matrices. As the G-matrices differed in their environment but not population, the genetic variances and covariances that contribute the most to the eigentensors are those which are most influenced by the environment. Eigentensor analysis was performed on the posterior median G-matrices. Uncertainty in the eigentensors was estimated by performing eigentensor analysis on the MCMC samples of the G-matrices. Finally, to determine whether an eigentensor explained more of the variation among G-matrices than would be expected by chance, we shuffled the real phenotypic data among environments, estimated G-matrices, and asked whether the eigentensors of the randomized G-matrices explained as much of the variation as the MCMC samples from the real data. If an eigentensor of the estimated G-matrices explain more of the variation, this indicates that this eigentensor is explaining biological variation and not only variation due to random sampling.

1.3 Results

We evaluated 7 phenotypes for each of 344 hybrids of doubled haploid (DH) lines crossed with a tester in replicated trials across 5 environments that varied in temperature, daylength, and watering or drought conditions (Supplemental fig. 1.5). Each DH line hybrid was genotyped for 551,460 SNPs, allowing us to identify ancestry segments along the genome.

1.3.1 Genotype x environment interactions

Genotypic main effects and GxE interactions contributed a significant amount of the variance of all measured traits (Fig. 1.1). Across environments, it was common for the rank of DH lines for grain yield to change, indicating that individual lines were generally not high yielding in all conditions (Fig. 1.1A). Anthesis-silking interval (ASI) showed a qualitatively similar pattern of rank-changing, while some traits such as thousand kernel weight (TKW) showed less dramatic GxE (Supplemental fig. 1.6). The proportion of variance due to main genotypic effects ranged from 0.34 for grain yield to 0.72 for male flowering date (Fig. 1.1B). For grain yield and HGM, GxE interactions contributed an amount of variance similar to the amount contributed by genotypic effects. For flowering time, TKW, and plant height, GxE interactions contributed less of the variance than main genotypic effects.

1.3.2 GWAS

Our test of the deviation effect of a marker within environments did not recover any markers significant at the 5% permutation threshold for any trait. In contrast, our plasticity GWAS identified two peaks which were significant at the 5% significance level, which were for ASI and female flowering (Fig. 1.2A, supplemental fig. 1.7). Neither of these peaks overlapped with GWAS peaks for main effects in this population (Odell *et al.*, 2022). The peak for ASI on chromosome 1 appears to be driven by the effect of the FV2 founder, which has a small effect in environments where ASI is close to zero but strongly increases the magnitude of ASI in environments where average ASI is greater (Fig. 1.2B). Patterns of identity by descent at the genomic region surrounding the peak identified unique haplotypes for 15 of the founders (Odell *et al.*, 2022), but a PCA of the SNPs in the region did not indicate that the FV2 haplotype was strongly diverged from other founders (Supplemental fig. 1.8). The peak for female flowering on chromosome 4 appears to be driven by founder A654, but the

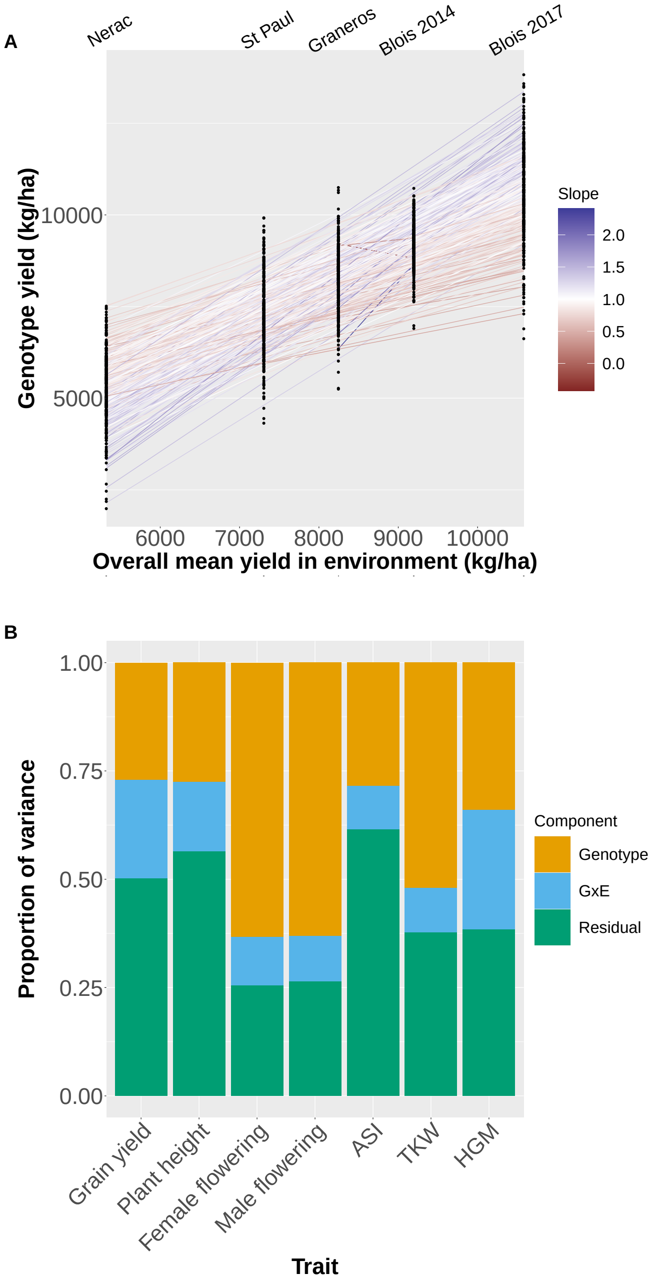

Figure 1.1: A) Mean yield of all genotypes in each environment. On the X axis environments are plotted by the mean yield across all genotypes in that environment. Points are mean yields of individual genotypes. Lines are the slope of a genotype's mean yield in each environment on the mean yield of all genotypes in that environment. The color of the line corresponds to the slope; a slope greater (or less) than one indicates a genotype more (or less) responsive to the environment than average. B) Restricted maximum likelihood estimates of variance components for each trait across all environments.

marker effects for this founder appeared unrealistically strong and likely reflect an artifact of the extremely low sampling of this founder among the DH lines. In addition to these two associations at the 5% level, we detected one peak which was significant at the 10% level for grain yield (Supplemental fig. 1.9). Our Finlay-Wilkinson GWAS uncovered one peak significant at the 5% level for ASI (Supplemental fig. 1.10). However, the founder whose effect appears to be driving this peak also appears to be underrepresented at this locus and only one line has a greater than 0.8 probability of carrying this founder allele. As a result, this peak is likely to be a statistical artifact.

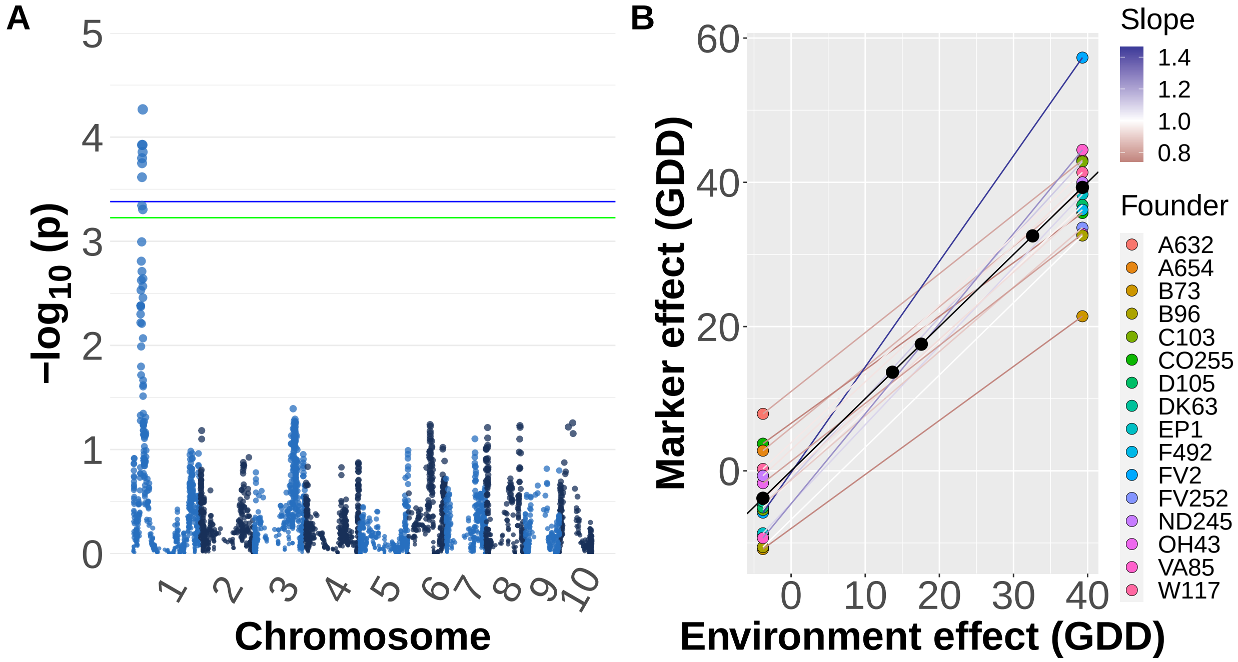

Figure 1.2: A) Manhattan plot for plasticity (model ii) GWAS on ASI. The blue and green lines represent the 5% and 10% significance levels based on permutation tests, respectively. B) Estimated effect of founder ancestry on plasticity for the most significant marker. The slope of a line indicates the plasticity of that haplotype and the difference in slopes is GxE. The color of the line corresponds to the slope; a slope greater (or less) than one indicates a genotype more (or less) responsive to the environment than average.

1.3.3 The G-matrix across environments

To understand how the environment affected pleiotropy, we estimated the genetic variance/covariance matrix (G-matrix) of five traits in each environment (Fig. 1.3A, B, supplemental figs. 1.11, 1.12). We dropped ASI and HGM from this analysis because models including those traits failed to converge; ASI was dropped due to concerns about collinearity as it is a function of two other traits in our analysis and HGM was dropped because in analyses run on subsets of these traits we found that HGM had very low covariance with the other traits. Comparisons of the 95% credible intervals of the difference between individual genetic correlations revealed numerous differences among environments (Supplemental fig. 1.13). Both the genetic variances of individual traits and the covariances between traits differed across environments (Fig. 1.3A, B). As the traits were mean scaled, the variances presented in Figure 1.3A are not heritabilites, which is the genetic variance scaled by the phenotypic variance. Importantly, mean-scaled genetic variances are not affected by the amount of residual variance, which means that a trait with high genetic variance relative to the mean along with high environmental variance can have low heritability but high mean-scaled genetic variance. (Houle, 1992). We found that grain yield generally had high mean-scaled genetic variance in each environment, and the single highest mean-scaled genetic variance of any trait in any environment was grain yield in Blois 2017. In one case, the sign of a genetic covariance changed: the genetic covariance between grain yield and female flowering date was positive across all environments except in Nerac 2016. This environment was the only one in which the values in the 2.5% and 97.5% quantiles of the posterior of the genetic covariance between grain yield and female flowering date was entirely negative, while in both years in Blois this interval was positive. The median posterior values of some other genetic covariances also switched signs between environments, but based on credible intervals we cannot state that they switched with confidence.

To quantitatively assess how individual elements of the G-matrix contributed to variation among environments, we performed an eigentensor analysis. The eigentensors of a set of G-matrices describe independent dimensions of variation among the G-matrices and can be used to identify which elements are contributing the most variation among the set. All of the four nonzero eigentensors explained significantly more variance than expected by chance (Supplemental fig. 1.14). The element of the G-matrix that most contributed to the first eigentensor was genetic variance for grain yield (Fig. 1.3C). When plotting each environment on this eigentensor, Blois 2017 is strongly differentiated from the other environments, which is probably due to the genetic variance for grain yield being the highest in this environment (Supplemental fig. 1.15). The genetic variance for grain yield also contributed strongly to the second eigentensor, while the genetic covariance between plant height and grain yield and the genetic variance of plant height contributed in the opposite direction. The third eigentensor described a contrast between genetic variance for plant height on the one hand and the genetic covariances between both female flowering date and TKW with grain yield on the other. Nerac is strongly differentiated on this eigentensor. While the covariance between female flowering and grain yield is not the only element of the G-matrix contributing to the third eigentensor, it is worth noting that Nerac is the only environment in which this covariance is negative.

Results of the analysis with non-scaled phenotypes are presented in the supplemental figures.

1.4 Discussion

1.4.1 Genotype x environment interactions

Genotype x environment interactions are known to be important for many agronomically important traits in maize, and our results on the relative importance of GxE across traits confirm these earlier findings. For example, male and female flowering date have been shown to be influenced predominantly by additive genetic effects and are not strongly influenced by GxE interactions (Buckler *et al.*, 2009; Rogers *et al.*, 2021), while grain yield and HGM have large GxE variance components relative to main genotype effects (Gage *et al.*, 2017; Rogers *et al.*, 2021). We find similar results in our analysis, indicating that this may be a consistent pattern for diverse maize germplasm in temperate environments.

If genotypes are adapted to different environments, we would expect to see GxE for fitness related traits. The high variance contributed by GxE to grain yield seen in this study thus indicates that the founder maize lines, despite all having been bred in temperate environments, still carry many alleles that are differentially adapted to this set of environments. For traits that are further removed from fitness it is less clear how to interpret the contribution of GxE. It may be that the GxE we observe for a trait like HGM, which has a high proportion of GxE variance and a low genetic covariance with grain yield, is an example of neutral plasticity and is not under strong selection (Des Marais *et al.*, 2013).

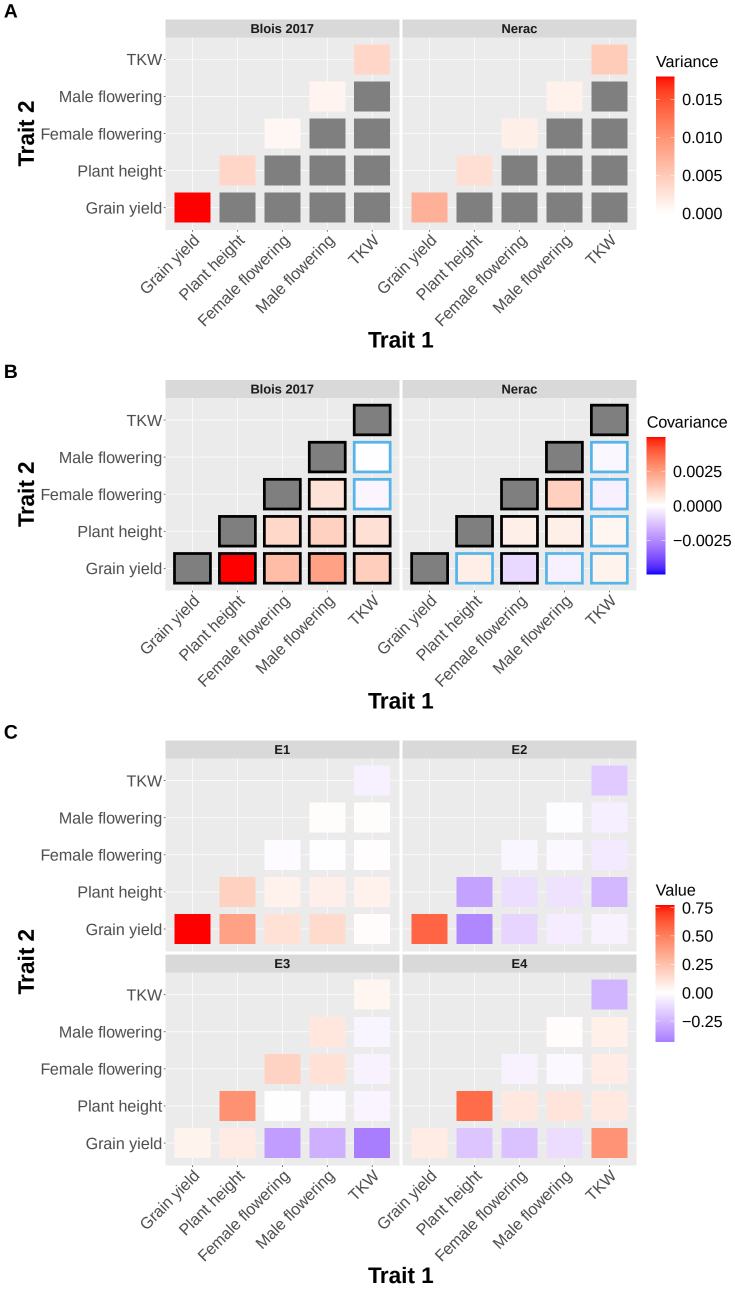

Figure 1.3: The genetic A) variances and B) covariances of the highest yielding environment (Blois 2017) and the lowest yielding environment (Nerac 2016). Traits are mean scaled. A black border around a covariance indicates that the 95% quantile interval of the posterior does not overlap with zero. Note that the scales on the upper and lower rows are different. C) Contributions of elements in the genetic variance-covariance matrices to the first four eigentensors of the set of genetic variance-covariance matrices. Elements on the diagonal are genetic variances of traits and elements on the off-diagonals are genetic covariances between traits. The color of a square represents the strength of the contribution of that element to the eigentensor, which is not dependent on the sign.

1.4.2 GWAS

Despite the presence of substantial GxE variance for several traits, we found relatively few markers which were significantly associated with GxE. One possible explanation is that the GxE variance we observed is largely polygenic and caused by many loci of small effect which we did not have power to detect with our GWAS. Previous studies investigating loci with main effects on traits such as grain yield and flowering time in maize have found that they are highly polygenic (Buckler *et al.*, 2009; Dell'Acqua *et al.*, 2015). It may not be surprising then if GxE for these traits also has a similarly polygenic basis. Grain yield is a highly integrated trait dependent on the interaction of many other traits with the environment; if those traits have a complex basis and different optima within different environments, then it would not be surprising to observe large GxE variance at the level of genotype while not observing significant GxE effects for individual loci.

1.4.3 The G-matrix across environments

The G-matrix has previously been shown to differ as much between environments as between populations (evidence reviewed in (Wood and Brodie, 2015)). Our work shows that the G-matrix differs across environments in a multiparent population of temperate maize lines. We find that these differences include both changes in the magnitude of genetic variances and covariances as well as changes in the sign of genetic covariances. The highest mean-scaled genetic variance we observed was for grain yield in Blois 2017, and in general grain yield had high mean-scaled genetic variance compared to other traits within each environment. This is in contrast to the finding that grain yield had the lowest heritability across all environments. This finding fits with previous work finding that fitness proximal traits frequently have low heritability but high mean-scaled genetic variance, possibly because of high residual variance for fitness proximal traits reducing heritability (Houle, 1992).

The magnitude of the genetic covariances between traits can be reduced solely as a function of reduced genetic variance for one or both of these traits without a change in the correlation between them. However, by looking at genetic correlations, we show that the correlations between traits varied across environments beyond effects of the differences in the variances (Supplemental fig. 1.13). Additionally, changes in the genetic variance alone will not cause the covariance between traits to change sign, which we also see for some combinations of traits. Particularly striking was the change in sign for the genetic covariance between grain yield and female flowering date observed in the most stressful environment, Nerac 2016. This environment was the only one in which the genetic covariance between grain yield and female flowering date was negative. Previous work has shown that flowering time is important for adaptation to drought stress (reviewed in Kazan and Lyons (2016)). Nerac 2016 experienced a drought from vegetative growth through maturity. Early flowering in this environment was genetically correlated with higher yields, suggesting that early flowering may have been a means to escape drought stress. The change in sign of the covariance is noteworthy given that we observed low GxE variance and high genotypic variance for female flowering date while simultaneously observing high levels of GxE variance for grain yield. This indicates that genotypes were relatively consistent in their flowering time across environments but that late flowering genotypes were higher yielding in most environments and lower yielding in one environment. In this way, a change in the genetic covariance between two traits (grain yield and female flowering) across environments may be contributing to GxE in one of those traits (grain yield), and provides an illustrative example of how traits that themselves show little GxE may nonetheless contribute to GxE for fitness.

While differences between environments presumably shape these changes in the G-matrix, previous work has found that neither measures of environmental novelty nor differences in phenotypic means predicted differences in the G-matrix when looking across all the studies in a meta-analysis (Wood and Brodie, 2015). In our analysis we find a similar result; differences between the G-matrices estimated in each environment are largely idiosyncratic and do not correspond with levels of stress or water availability. Eigentensor analysis reveals that each of the main directions of variation across G-matrices correspond mostly to the differentiation of one or at most two of the environmental G-matrices from the others. Previous work investigating the G-matrix of plant populations grown in well-watered and drought environments has been inconsistent in terms of whether drought stress increases or decreases genetic variance and how it affects the genetic correlation between flowering time and yield (Manzaneda *et al.*, 2015; Sherrard *et al.*, 2009). Considering our work in the context of previous studies, we suggest that the environmental contribution to the G-matrix is complex and not easily described by one environmental axis, which raises the possibility that multivariate adaptation to the environment may be difficult to predict.

Additionally, both the severity and timing of drought seem to be important in determining the effects of water deficit on covariances between traits. In this study we find that in Nerac, the most drought stressed environment, the genetic covariance between flowering time and yield is negative and that this genetic covariance contributes to differentiating it from the other environments. The fact that the genetic covariance between flowering date and grain yield in the other water deficit environment, St. Paul, was not significantly negative may be because that population was given water during flowering while in Nerac water deficit extended through flowering. It appears that how the G-matrix is affected by environmental stress is highly dependent on the species and population studied and the exact stress applied.

1.5 Conclusion

Using a MAGIC population of maize grown in five environment x year combinations we were able to analyze the genetic basis of GxE in a set of diverse maize lines. We observed GxE variance for all traits and for some traits we observed comparable amounts of genotypic and GxE variance. Estimating the G-matrix within each environment revealed that changes in genetic variances and covariances across environments were common. Notably, the genetic covariance between yield and female flowering time was positive in most environments but negative in one of the environments. GWAS identified one locus significantly associated with GxE for anthesis-silking interval. Given the substantial GxE variance, the low number of significant loci suggests that GxE for the traits we analyzed may have a polygenic basis.

1.6 Acknowledgments

We want to thank Silas Tittes and Sarah Turner-Hissong for their advice on statistical analysis. A.I.H. was supported by National Science Foundation Graduate Research Fellowship 1650042 and the UC Davis Dept. of Plant Sciences. J.R.I. was supported by USDA Hatch project CA-D-PLS-2066-H 548. S.O. was supported by the UC Davis Dept. of Plant Sciences and NSF grant 1754098. D.E.R. was supported by USDA Hatch project 1010469 and USDA NIFA grant no. 2020-67013-30904. This work was previously published as Hudson *et al.* (2022) and we thank Jean-Luc Jannink for editing and two anonymous reviewers for their feedback.

1.7 Appendix

1.7.1 Supplementary figures and tables

Table 1.1: **Features of the five growing environments.**

Environment-Year	Mean temperature (°C)	Mean relative humidity (%)	Mean precipitation (mm)	Water treatment[1]	Planting density (seeds/hectare)
Blois 2014	16.7	75.2	2.19	OPT	85,000
Blois 2017	17.0	72.3	1.71	OPT	95,000
Graneros 2015	20.1	55.1	0.266	OPT	90,000
Nerac 2016	19.1	74.9	1.15	Early term	85,000
St. Paul 2017	20.3	65.4	1.12	Recovery	90,000

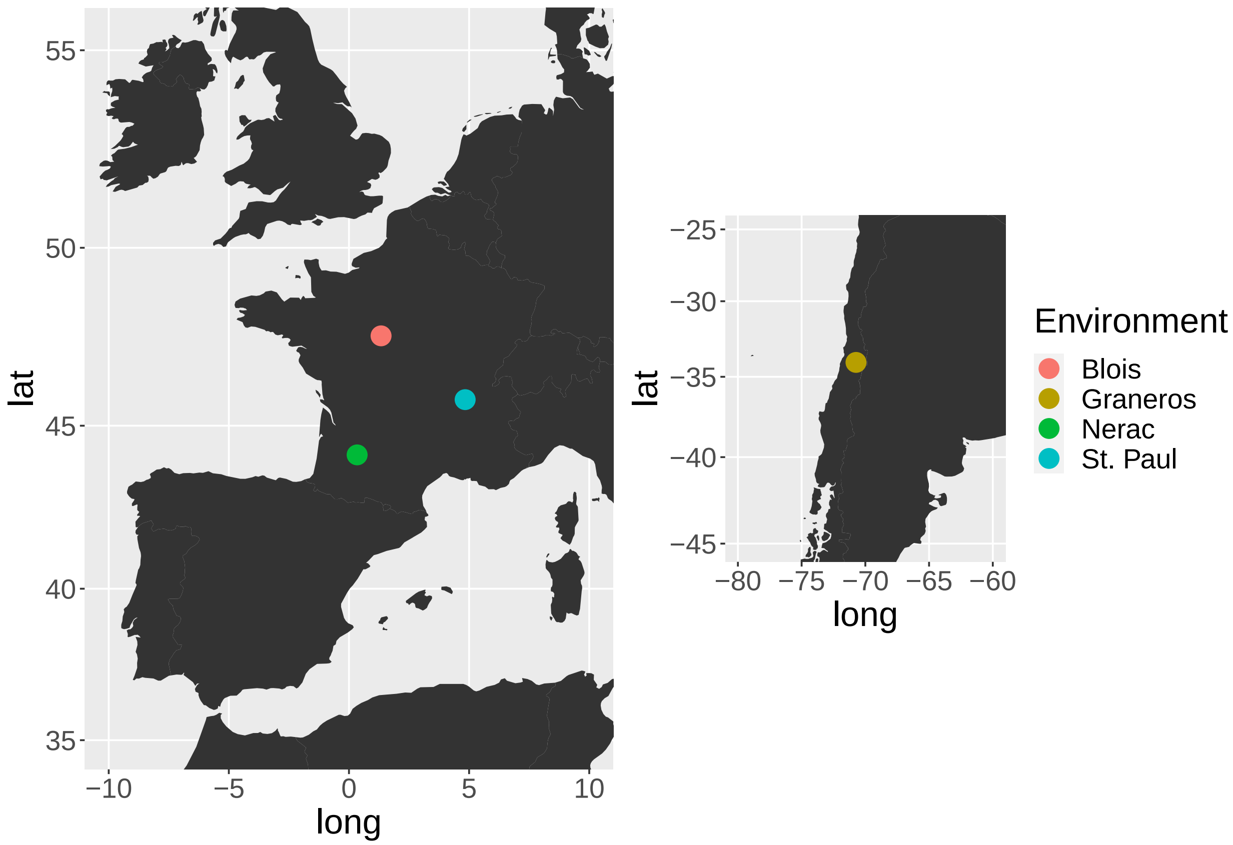

Figure 1.4: Locations of the environments the MAGIC population was grown in. In one environment (Blois) the MAGIC population was grown in two years.

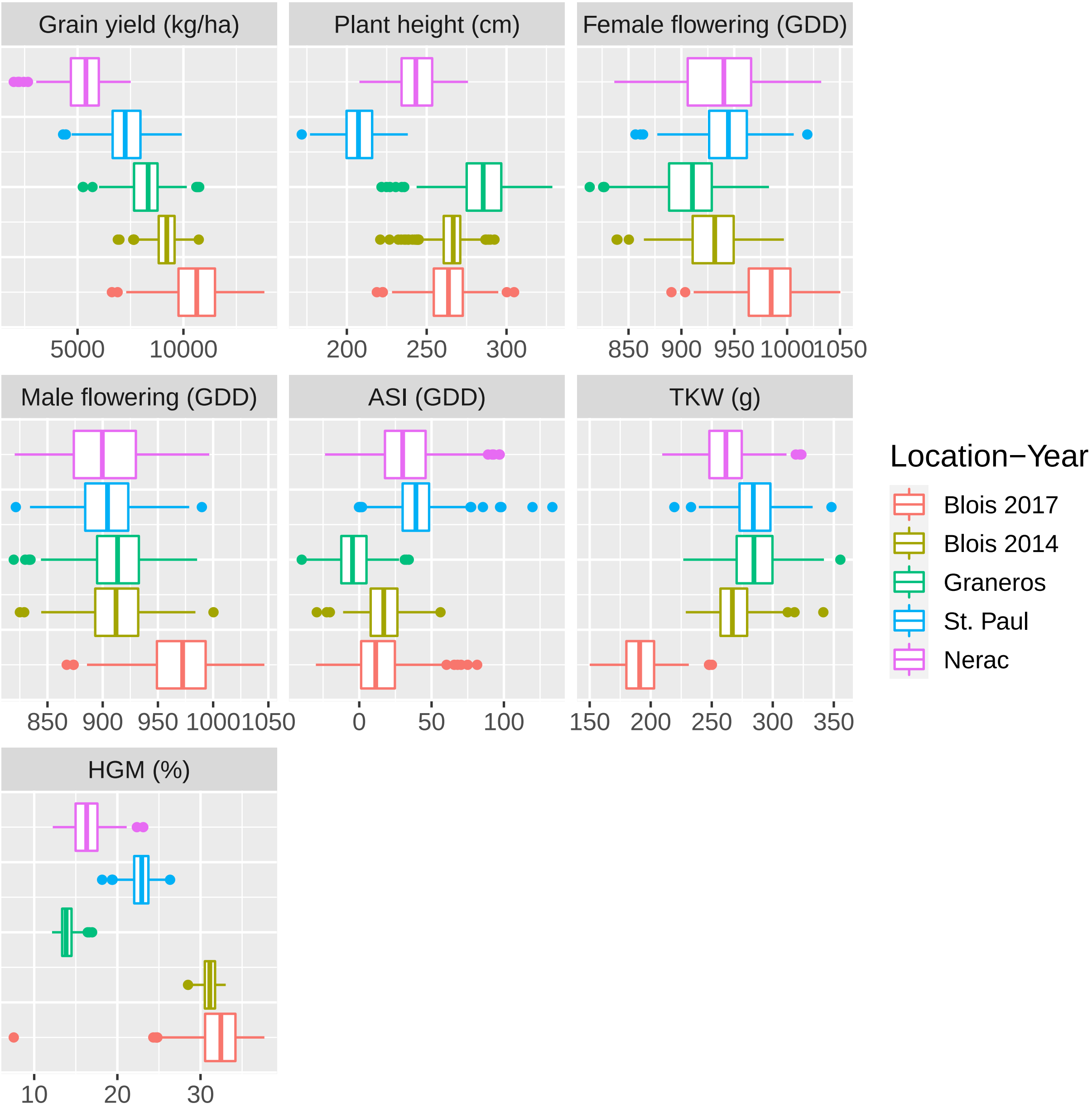

Figure 1.5: Phenotypes of the MAGIC population in each environment. Box plots show median as line in box, 25% and 75% quartiles as edges of box, and most extreme values that are at most 1.5 * inter-quartile range as whiskers, with any values more extreme shown as individual dots.

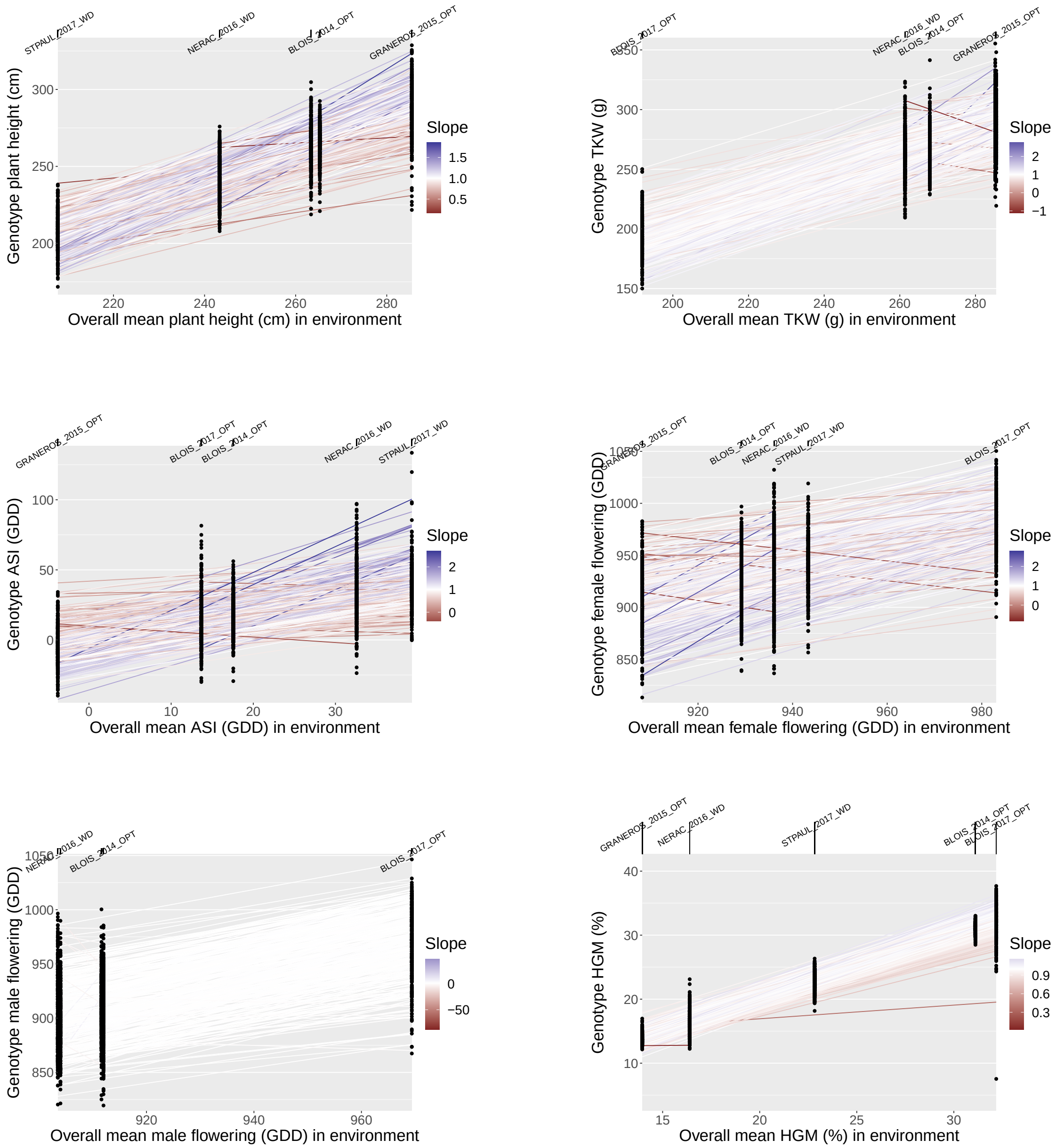

Figure 1.6: Mean trait values of all genotypes in each environment. On the X axis environments are plotted by the mean of each trait across all genotypes in that environment. Circles are the mean trait values of individual genotypes. Lines are the slope of a genotype's mean trait value in each environment on the mean trait value of all genotypes in that environment. The color of the line corresponds to the slope; a slope greater (or less) than one indicates a genotype more (or less) responsive to the environment than average.

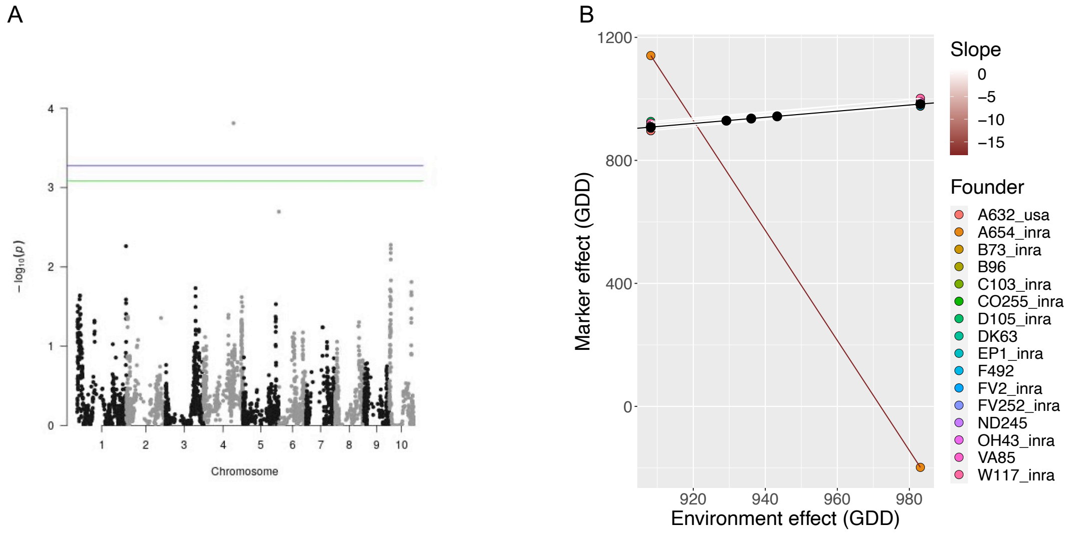

Figure 1.7: A) Manhattan plot for plasticity GWAS on female flowering. The blue and green lines represent the 5% and 10% significance levels based on permutation tests, respectively. B) Estimated effect of founder ancestry on plasticity for the most significant marker. Lines are the slope of a marker's effect in each environment on the mean female flowering date of all genotypes in that environment. The color of the line corresponds to the slope; a slope of one indicates a marker with the average response to the environment, a slope less than one indicates a marker less responsive to the environment than average, and a slope greater than one indicates a marker more responsive to the environment than average. Effect sizes in GDD.

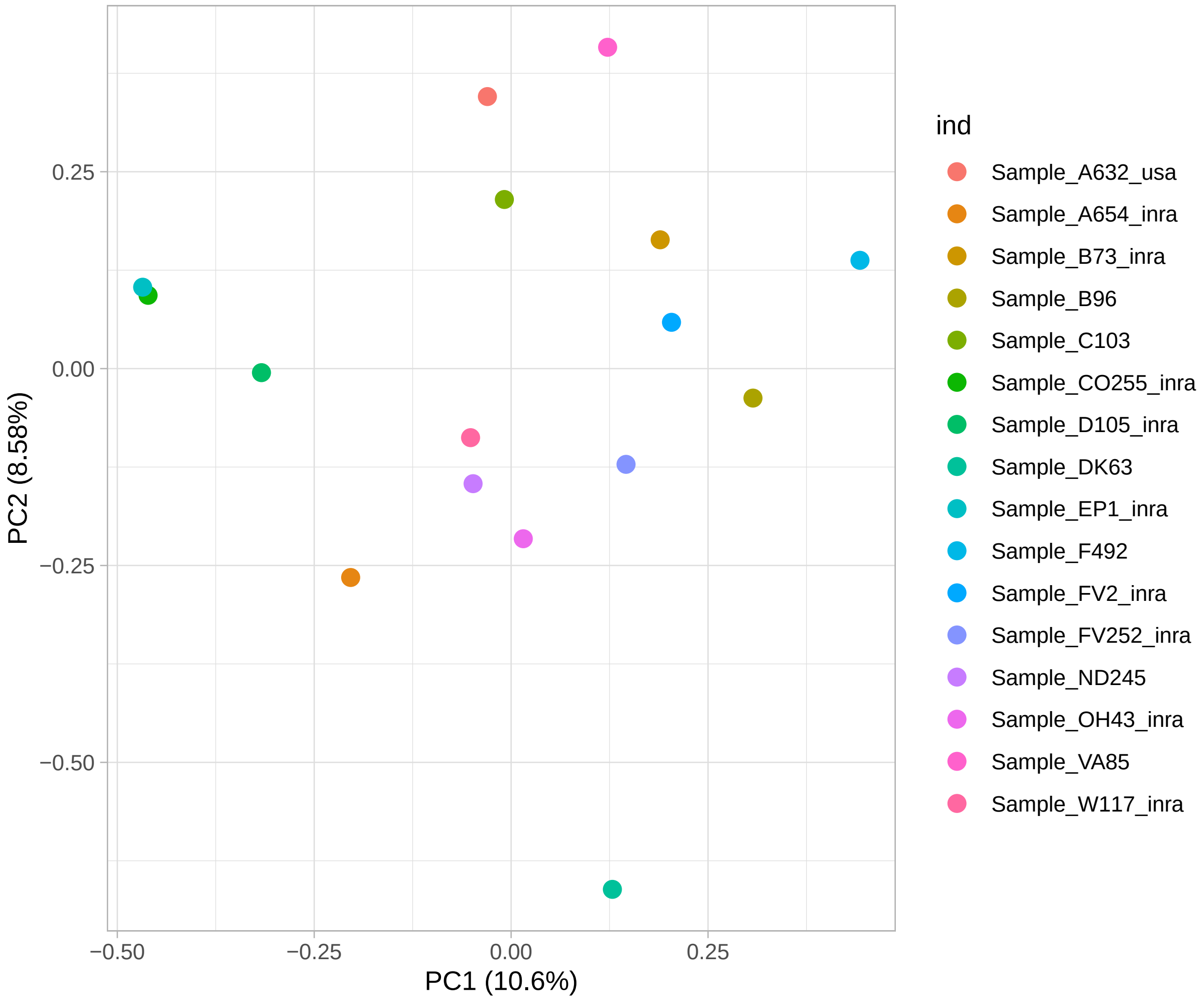

Figure 1.8: Each founder plotted on the first and second principal components from a principal component analysis of the SNPs within the plasticity GWAS peak for ASI.

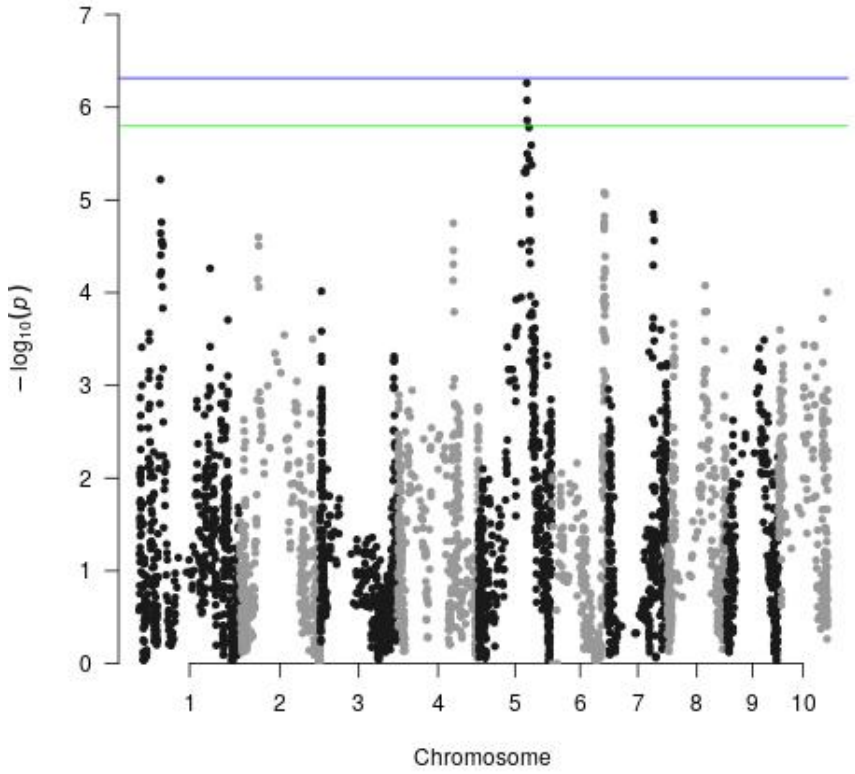

Figure 1.9: Manhattan plot for plasticity GWAS on grain yield. The blue and green lines represent the 5% and 10% significance levels based on permutation tests, respectively.

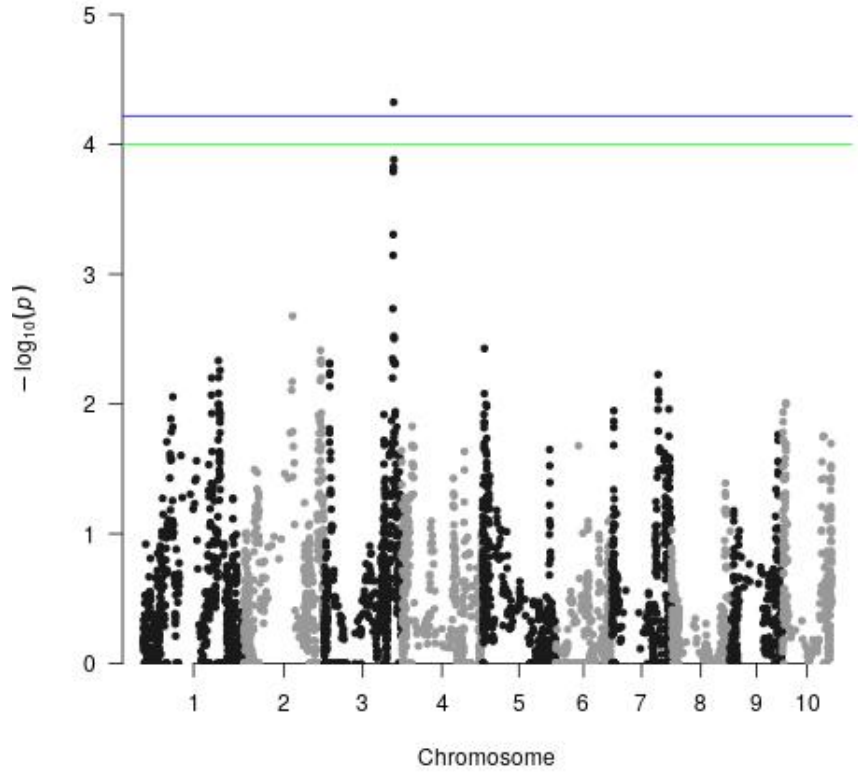

Figure 1.10: Manhattan plot for Finlay-Wilkinson GWAS on ASI. The blue and green lines represent the 5% and 10% significance levels based on permutation tests, respectively.

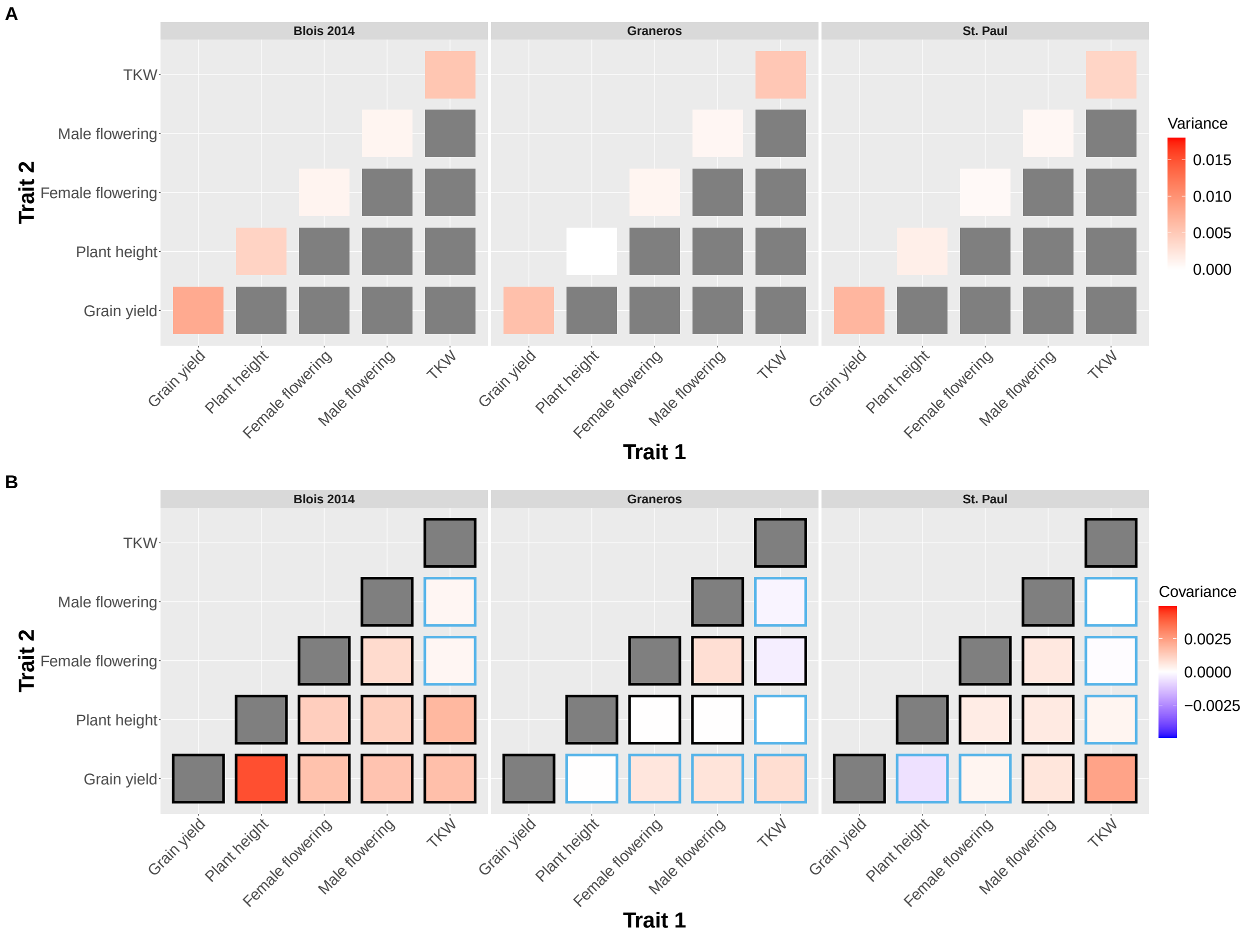

Figure 1.11: Heat maps of the G-matrices for the remaining environments. Genetic variances are on the top row (A) and covariances are on the bottom row (B). A black border around a covariance indicates that the 95% quantile interval of the posterior does not overlap with zero.

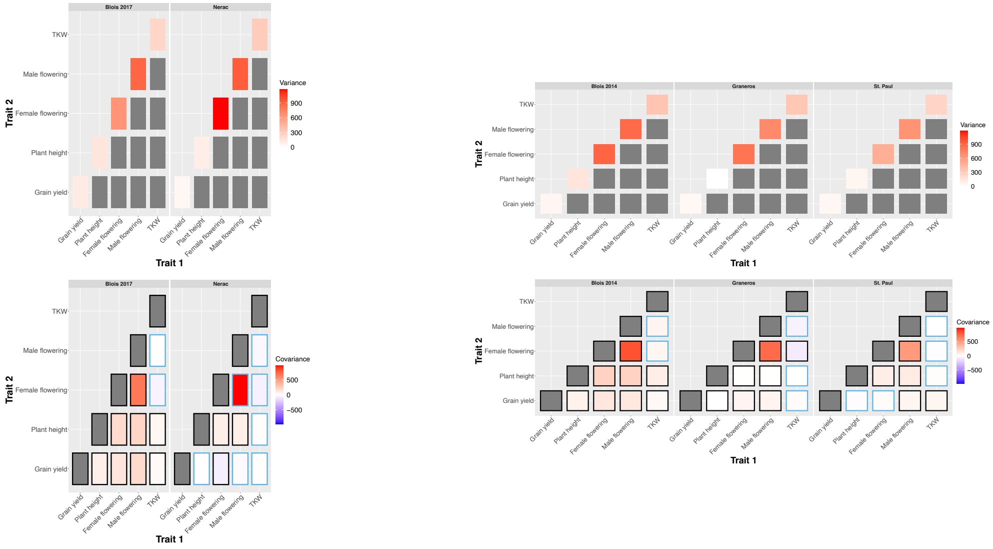

Figure 1.12: Heat maps of the G-matrices with phenotypes not scaled. Note that this means all traits are on different units. Grain yield is in quintals/ha, plant height is in centimeters, female flowering and male flowering are in GDD, and TKW is is grams. A black border around a covariance indicates that the 95% quantile interval of the posterior does not overlap with zero.

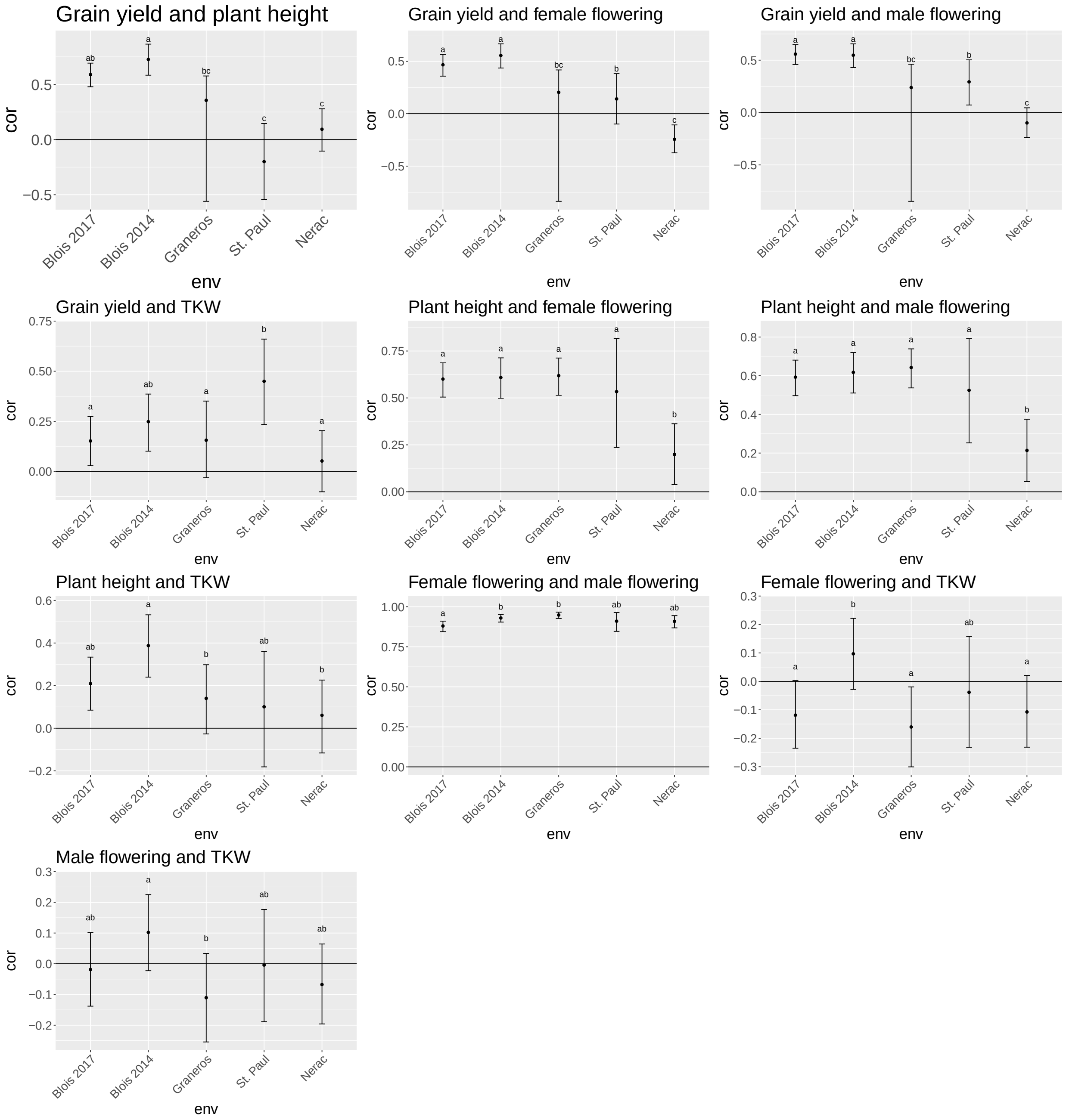

Figure 1.13: Genetic correlations of each pair of traits. For each pair of traits genetic correlations are shown for each environment with 95% credible intervals. Letters indicate significantly different groups as determined by comparing the 95% credible intervals of the difference between MCMC samples from estimating the correlation in each environment.

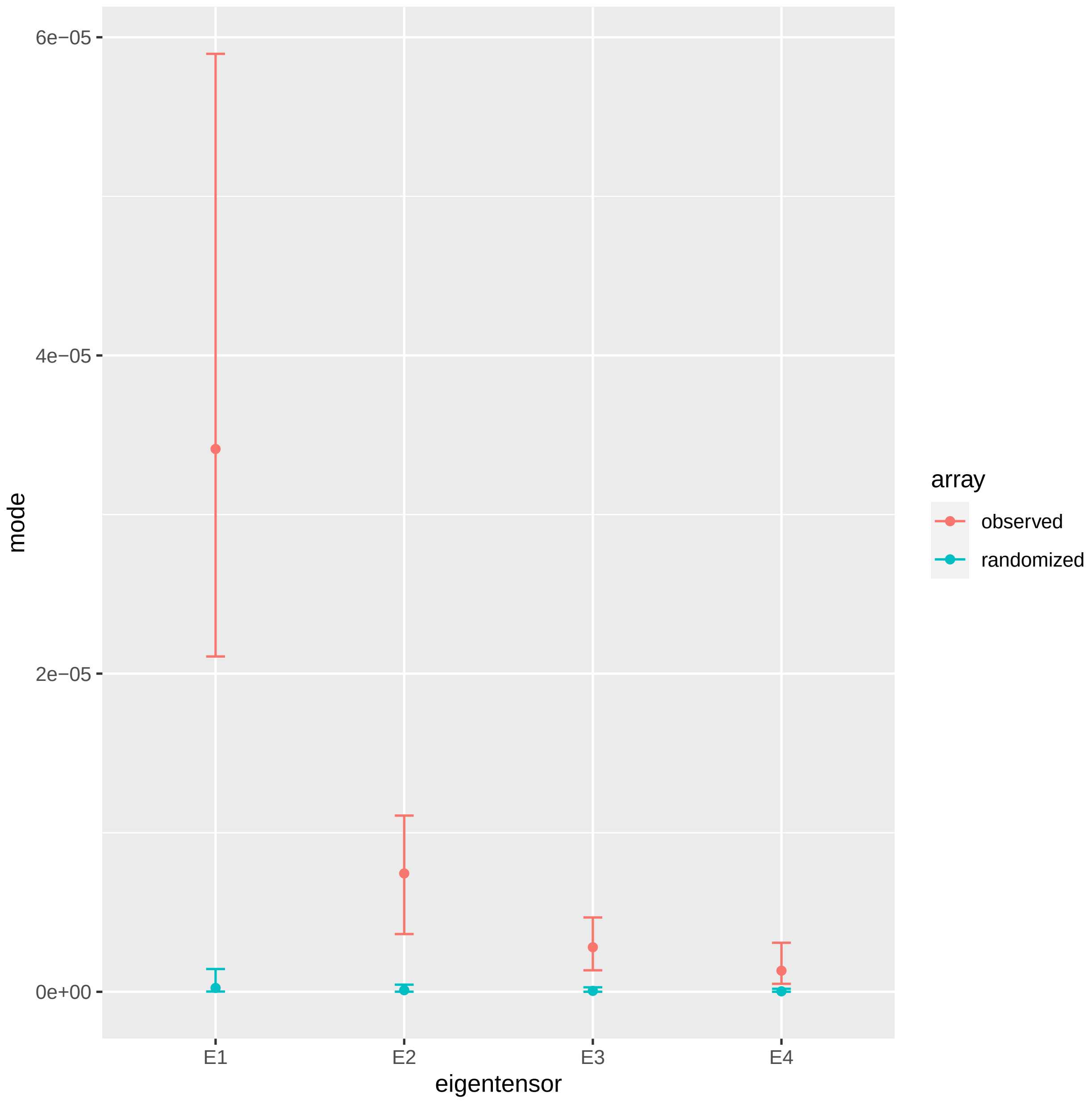

Figure 1.14: Posterior mode and 95% credible intervals of the eigenvalues of the non-zero eigentensors of the G-matrices. If the observed eigenvalue of an eigentensor is greater than the 95% credible intervals of eigenvalues of the eigentensors estimated from randomized data that indicates the eigentensor explains more of the variation among G-matrices than would be expected by chance.

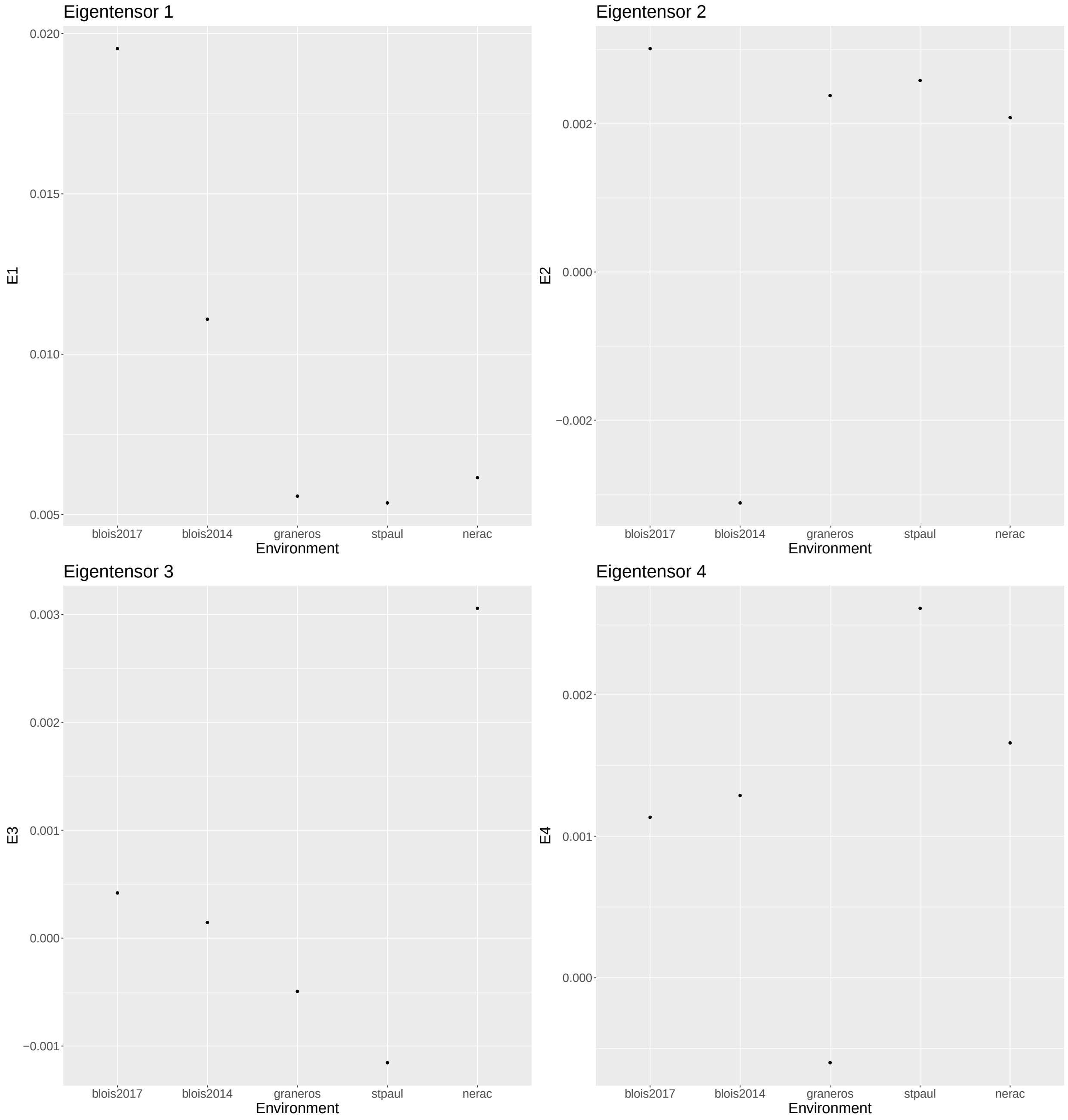

Figure 1.15: The G-matrix estimated in each environment plotted on each of the four first eigentensors. Note that the scale on the y axis is different for each plot.

Chapter 2

Structural variants are depleted in constrained regions of the maize genome

Asher I. Hudson[1,2], Arun Seetharam[3,4], Matthew B. Hufford[3] Jeffrey Ross-Ibarra[1,2,5]

[1] Department of Evolution and Ecology, University of California, Davis, CA, USA

[2] Center for Population Biology, University of California, Davis, CA, USA

[3] Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, USA

[4] Genome Informatics Facility, Iowa State University, Ames, IA, USA

[5] Genome Center, University of California, Davis, CA, USA

2.1 Introduction

Structural variants (SVs) are common in the genomes of many species (Feuk *et al.*, 2006; Fuentes *et al.*, 2019; Yang *et al.*, 2019). Compared to single nucleotide polymorphisms (SNPs), however, SVs are relatively less studied and understood. This is in large part because SVs are more difficult to identify in genomic data than SNPs (Mahmoud *et al.*, 2019). Despite this, there are numerous examples of SVs with phenotypic effects, including several implicated in adaptation and domestication (Shomura *et al.*, 2008; Zancolli *et al.*, 2019; Su *et al.*, 2019).

While some SVs are adaptive, many are likely deleterious. Overall, theoretical and empirical evidence suggests that most mutations with phenotypic effects are deleterious (Joseph and Hall, 2004; Keightley and Lynch, 2003). Different types of mutations can have different distributions of fitness effects, which is clear in SNPs as non-synonymous and biochemically non-conservative SNPs are on average more deleterious than synonymous and conservative SNPs, respectively (Cargill *et al.*, 1999). Similarly, SVs may on average be more or less deleterious than SNPs. Different types of SVs, such as insertions and deletions, may also differ from each other. SVs can also differ from each other based on size, whether the size indicates the amount of sequence inserted or the number of base pairs inverted.

In this study, we analyze the distribution of SVs in the maize genome using genome assemblies of 25 maize inbred lines. These assemblies were created with a hybrid long read-short read approach and optical mapping, allowing us to analyze many more SVs than previously possible in this species. Maize is an ideal system for studying structural variation as it is known to be abundant in the maize genome (Schnable *et al.*, 2009; Brunner *et al.*, 2005). 83.20% of the genome of the maize line B73 is estimated to be composed of transposable elements (TEs), which are one cause of presence absence variation and have been associated with agriculturally important traits (Hufford *et al.*, 2021; Yang *et al.*, 2013). Two examples of functionally important structural variation in maize are a 14 Mb inversion on chromosome 4 associated with earlier flowering time and that may be adaptive in highland maize and deletions in the gene *Waxy* that are associated with the waxy phenotype (Pyhäjärvi *et al.*, 2013; Romero Navarro *et al.*, 2017; Okagaki *et al.*, 1991; Fan *et al.*, 2008). If most new mutations are deleterious, however, examples of adaptive SVs may be the exception rather than the rule.

Here we identify evolutionarily constrained regions in the maize genome and then ask whether SVs are enriched or depleted in these regions. We find that SVs are strongly depleted overall in constrained regions and that they are more depleted than SNPs, suggesting that on average they may be more deleterious. This pattern varies greatly among classes of SVs and we identify genomic features that explain some of these differences.

2.2 Methods

2.2.1 GERP

We identified conserved regions of the maize genome using a comparative genomic approach. To do this, we obtained 13 publicly available angiosperm genomes from Ensembl Plants and Phytozome. Soft masked copies of the genomes were aligned to the unmasked B73 v5 reference genome using Last (Kiełbasa *et al.*, 2011). Repetitive elements in B73 v5 were then masked in the aligned sequences using a RepeatMasker file. We identified a species tree using previously published trees (Soreng *et al.*, 2017; Smith and Brown, 2018). We then estimated neutral evolutionary rates for branches of the tree from fourfold degenerate sites in the alignment using rphast (Hubisz *et al.*, 2018). We used the tools gerpcol and gerpelem from GERP++ to estimate conservation scores at aligned base pairs and identify conserved elements (Davydov *et al.*, 2010a). GERP++ conservation scores are based on rejected substitutions across species, where sites that have fewer substitutions than expected based on the neutral tree are given more positive scores. We did not include the maize genome when calculating GERP scores in order to avoid reference bias.

2.2.2 Enrichment analyses

Bed files with SVs are from Hufford *et al.* (2021). Briefly, Hufford et al. sequenced and assembled 26 inbred maize genomes with a hybrid approach using PacBio long read sequencing, Illumina short reads, and Bionano optical mapping, allowing a higher degree of contiguity and completeness than previous assemblies. This approach was also able to assemble large amounts of the repetitive regions of the genome. They then identified SVs by aligning both long reads and whole genome assemblies of the other 25 inbreds to B73. Importantly, this means that all SVs are relative to the B73 genome. It is difficult to determine whether a given insertion deletion polymorphism (indel) is an insertion or deletion relative to the ancestral state. As our GERP elements are also called based on alignments to B73, we use the same polarization for insertions and deletions as well as using combined indels for analyses.

To test whether SVs were depleted in conserved elements, we measured the overlap between SVs and conserved elements and performed Fisher's exact tests. For tests involving combined deletions and insertions, we measured the overlap of base pairs in conserved elements with the presence of a SV in any of the NAM parental lines. We also tested for the depletion of deletions and insertions in conserved coding sequence, conserved noncoding sequence, and conserved non-genic sequence. In all three of these cases, the Fisher's exact test was testing depletion compared with non-conserved elements. For tests involving insertions, we also measured the overlap of GERP elements with insertion start sites. As insertions may simply move conserved elements while maintaining their function, we speculated that insertion start sites may be more meaningful than base pairs of overlap with conserved elements (Fig. 2.1). Insertions were also subdivided into quartiles based on size to test whether the size of insertions was associated with depletion in GERP elements.

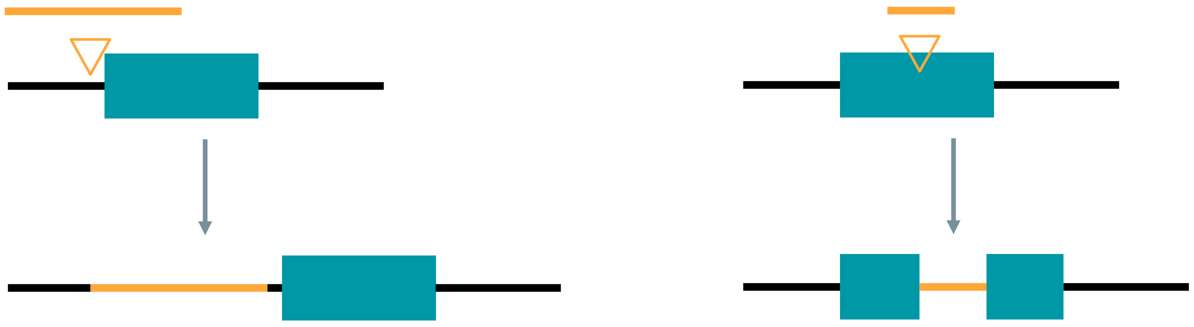

Figure 2.1: Illustration of potential impacts of an insertion. On the left, an insertion shifts the location of a conserved element without interrupting it. On the right, an insertion interrupts the conserved element.

2.3 Results

2.3.1 Enrichment analyses

Both SVs and SNPs were significantly depleted in conserved elements based on Fisher's exact tests, but SVs were depleted to a greater degree (Fig. 2.2). The log odds ratios for both SVs and SNPs were increased when considering conserved elements outside of coding sequence and outside of genic sequence and slightly reduced when considering conserved elements in coding sequence. All log odds ratios were significantly lower than zero. A log odds ratio lower than zero indicates depletion, with a lower log odds ratio indicating greater depletion. When we partitioned SVs by type, we found striking differences in their depletion in conserved elements (Fig. 2.3). Combined indels were significantly depleted in conserved elements. Deletions by themselves were even more significantly depleted in conserved elements. Insertions were actually enriched in conserved sequence, contrary to our expectations. Translocations were also significantly depleted, although to a much lesser degree. Inversions were not significantly depleted.

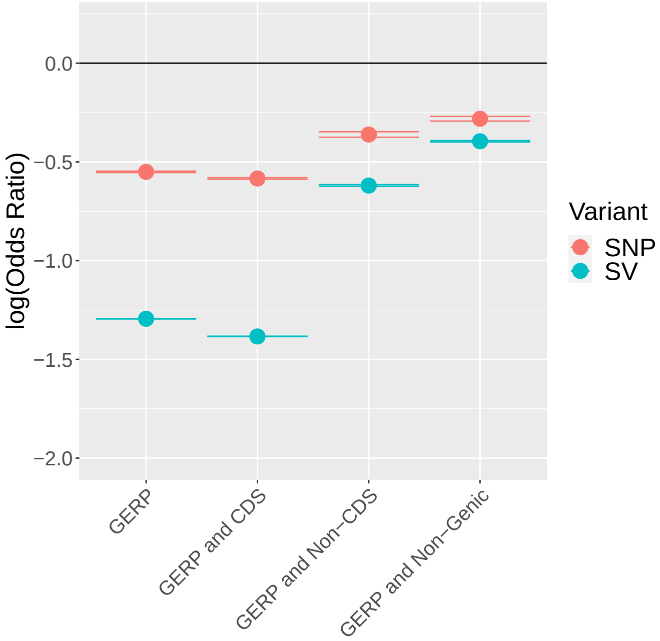

Figure 2.2: The log ratio of the odds that an allele overlapped with a bp in a conserved element versus outside the conserved elements. Error bars are 95% confidence intervals. Conserved elements are GERP elements, the intersection of GERP elements and coding sequence, the intersection of GERP elements and non-coding sequence, and the intersection of GERP elements and non-genic sequence.

To further investigate the surprising enrichment of insertions in conserved elements, we performed a logistic regression of proportion of insertion base pairs in a 10 kb window on number of base pairs in conserved elements, recombination rate, open chromatin, and number of base pairs masked in the B73 reference (Fig. 2.4). The number of conserved base

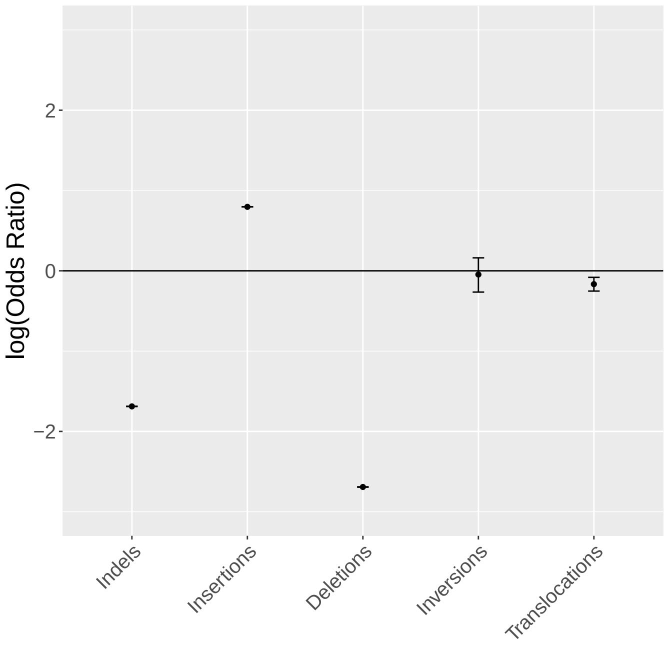

Figure 2.3: The log ratio of the odds that a base pair overlapped a SV given that it was in a conserved element versus it being outside the conserved elements. SVs are partitioned by type. Error bars are 95% confidence intervals.

pairs had a significant negative effect on the proportion of insertion base pairs. The discrepancy between this result and the apparent enrichment of insertions in conserved elements is due to the correlation between conserved elements and the other features in our regression model. Insertions are positively correlated with both higher recombination rates and open chromatin, which are themselves positively correlated with constrained base pairs. Additionally, conserved elements are strongly negatively correlated with masked sequence. As masked sequence has a negative effect on proportion of insertion base pairs, this contributes to the enrichment of insertions in conserved elements.

The start sites of all size quartiles of insertions were significantly depleted in conserved elements (Fig. 2.5). The largest insertions were more depleted in conserved elements than small and medium sized insertions. The depletion of insertion start sites may seem perplexing given the apparent enrichment of insertion base pairs, but is explained by the fact that most conserved elements either entirely overlap with insertions or do not overlap at all. This means that while there are many conserved elements that are moved by insertions relative to their positions in B73, there are very few that are interrupted.

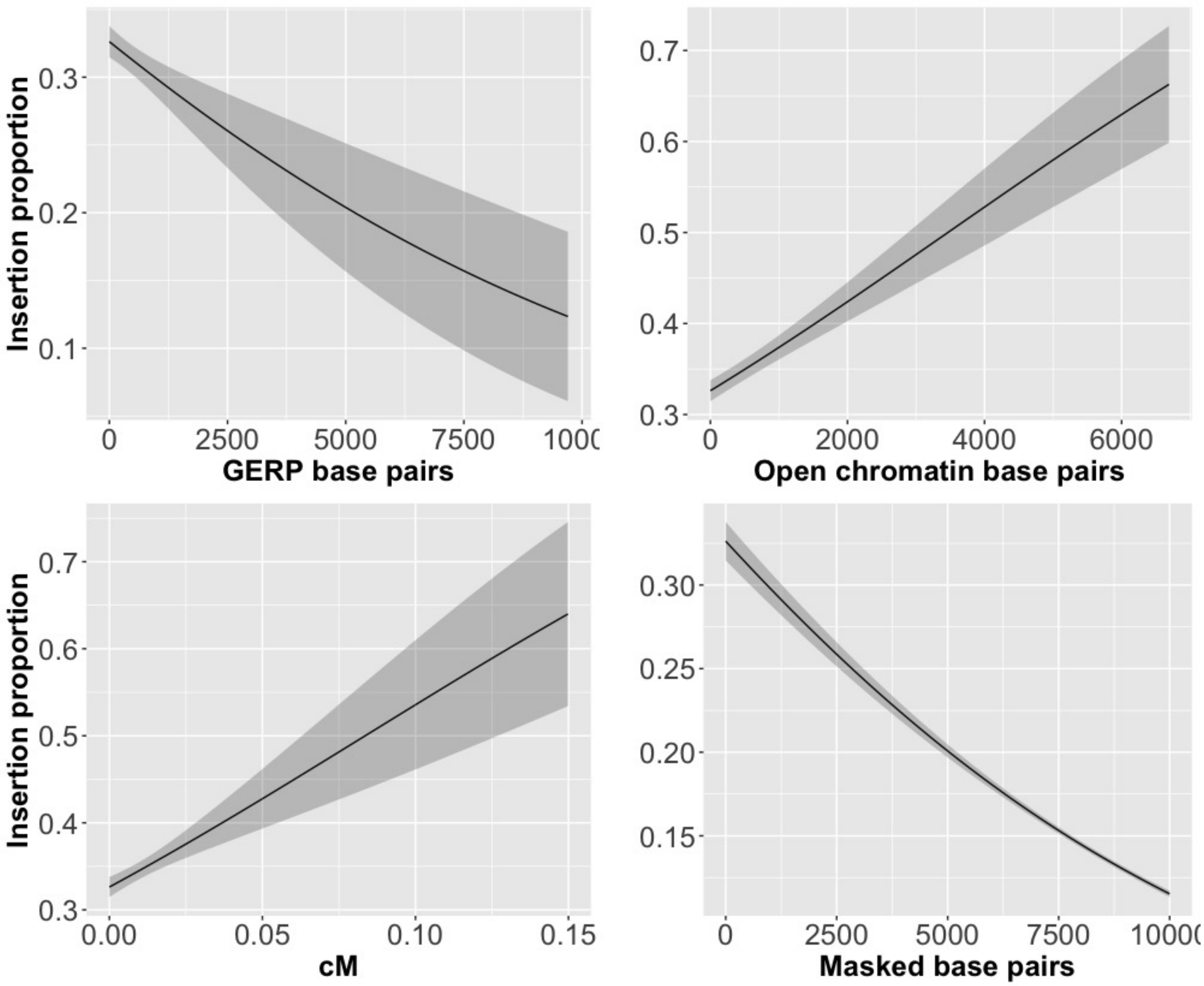

Figure 2.4: Most 10 kb windows have no insertion base pairs, but smaller numbers have proportions up to 1.0. Recombination rate and number of open chromatin base pairs in a window are positively associated with the proportion of a window covered by insertions, while the number of conserved element base pairs and masked base pairs are negatively associated.

2.4 Discussion

We found that SVs are strongly depleted in constrained regions of the genome. Additionally, we found that SVs are more depleted in these regions than SNPs. While this pattern is stronger in constrained regions within coding sequence, it is also visible in constrained regions both outside of coding sequence and outside of genes.

While SVs are depleted overall, this effect varies by class. Deletions are strongly depleted in constrained regions. Inversions and translocations appear to be neither strongly enriched

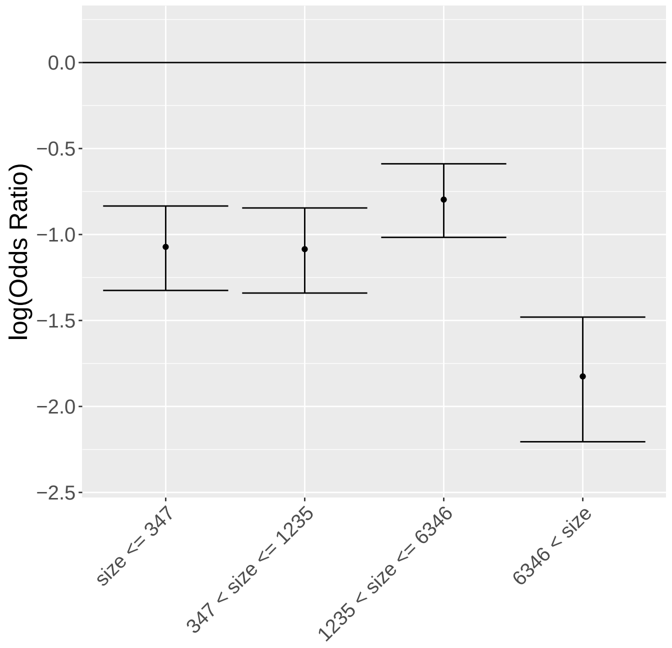

Figure 2.5: The log ratio of the odds that a base pair overlapped the start site of an insertion given that it was in a conserved element versus it being outside the conserved elements. Error bars are 95% confidence intervals. Insertions are binned according to quartiles of size.

nor strongly depleted. Unlike all other types of SVs, insertions were enriched in constrained regions. While we were surprised that insertions appeared to be enriched, the differences in enrichment/depletion between classes of SVs are in line with previous research. In *Drosophila melanogaster*, deletions occur more often than insertions but are also more likely to be under negative selection and less likely to be fixed by positive selection (Leushkin *et al.*, 2013). Inversions don't interrupt genes or regulatory elements unless those elements overlap with the inversion breakpoints. As a result, unless gene expression is altered, inversions may not in themselves have large effects on phenotype, and there is evidence that inversions do not generally alter gene expression in *Drosophila* (Ghavi-Helm *et al.*, 2019). Inversions do disrupt normal recombination, however, and can accumulate deleterious alleles as well as combinations of adaptive alleles (Kirkpatrick, 2010). It may not be surprising then that inversions do not appear to be strongly depleted, especially when compared to the more unambiguous evidence for new deletions being likely to be deleterious.

The unexpected enrichment of insertions in constrained elements lead us to investigate other factors that might explain this pattern. We found that this enrichment is partially explained by insertions being positively correlated with recombination and open chromatin. There is evidence that recombination can be mutagenic in several species (Yang *et al.*, 2015). Some classes of TEs, such as Mu elements, are more likely to insert in gene-rich, high recombination regions (Schnable *et al.*, 2009). Thus, a higher rate of insertion generating mutations may be masking selection against insertions in these regions. Recombination is also correlated with double-strand breaks, repair of which can result in indels (Lieber, 2010). This may also be true for inversions, which can be caused both by repair of double-strand breaks and ectopic recombination (Huang and Rieseberg, 2020). Additionally, insertions were negatively correlated with masked base pairs. The negative effect of masked sequence on insertion proportion is likely due to the technical issue of insertions being more difficult to call in regions with large amounts of repetitive sequence. This suggests that our enrichment analyses are likely to be in general under-estimating the depletion of SVs in conserved elements. Finally, we found that insertions which interrupt constrained regions, rather than shifting their coordinates within the genome, are relatively rare, suggesting that moving constrained elements may not be strongly selected against.

As mentioned previously, it is difficult to determine whether indels are insertions or deletions relative to the ancestral genome, and this makes interpretation of results with these SVs complicated. In conserved elements it may be more likely that we are able to correctly infer the type of mutation that occurred. If an element is conserved across multiple species, this is likely to represent the ancestral sequence. A deletion within this element in B73 may also be a deletion relative to the ancestral sequence, and the same for an insertion. However, this does not solved the problem of polarizing indels by ancestral state outside of conserved elements, which is also critical for estimating enrichment. Additionally, it is possible that some polymorphisms that we call insertions in lines other than B73 may actually be evolutionarily conserved elements that have been deleted in B73. This could potentially contribute to the apparent enrichment of insertions that we observed. Some of our results which polarize indels relative to B73 may be robust to this issue, such as the finding that large insertions within conserved elements are more depleted than smaller insertions. In order to fully understand selection on insertions and deletions separately further work with mutation accumulation lines and ancestral genome reconstruction may be informative.

SV abundance is negatively correlated with evolutionary constraint when accounting for other genomic features, suggesting that most SVs are under negative selection. Experiments to determine mutation rates for different types of SVs along the genome could help to clarify the relative contributions of selection and mutation to the distribution of SVs and to provide further clarification on which types of mutations are more likely to be deleterious.

2.5 Acknowledgements

We want to thank Arun Seetharam, Matthew Hufford, and Margaret Woodhouse for providing data on structural variants and syntenic genes. A.I.H. was supported by National Science Foundation Graduate Research Fellowship 1650042 and the UC Davis Dept. of Plant Sciences. Portions of this work were previously published in Hufford *et al.* (2021).

Chapter 3

Identifying evolutionarily constrained regions in the maize genome using machine learning

Asher I. Hudson[1,2], Ashley Johnson[1], Arun Seetharam[3,4], Jeffrey Ross-Ibarra[1,2,5]

[1] Department of Evolution and Ecology, University of California, Davis, CA, USA

[2] Center for Population Biology, University of California, Davis, CA USA

[3] Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, USA

[4] Genome Informatics Facility, Iowa State University, Ames, IA, USA

[5] Genome Center, University of California, Davis, CA, USA

3.1 Introduction

Eukaryotic genomes are a mixture of genic and non-genic sequence. While some of this non-genic sequence is known to be regulatory, what fraction of the genome is functional is still under debate (Schrider and Kern, 2015; Graur *et al.*, 2013; Davydov *et al.*, 2010b; The ENCODE Project Consortium, 2012). Regions of the genome that are functional are likely to be evolutionarily constrained as mutations are more likely to have deleterious effects on fitness in these regions. Multiple important questions in evolutionary biology are directly related to understanding which regions of the genome are constrained. The fraction which affects phenotype limits the target size for adaptive evolution. It also limits the target size for deleterious mutations, as mutations outside of that fraction should have little or no effect on fitness.

The fraction of the genome experiencing selection is almost certain to vary among species due to a variety of factors. The maize genome offers a chance to test several questions about which regions of the genome are evolutionarily constrained. As in many plants, synteny is relatively low and sequences outside of coding sequence evolve rapidly, making it difficult to identify which regulatory sequences might be important (Zhao and Schranz, 2019). Maize is also an ancestral allotetraploid; the ancestor of maize was a hybrid between two species which subsequently reduced to a diploid with genomic regions inherited from both parental species (Gaut and Doebley, 1997). While a process of gene fractionation has resulted in the loss of one copy of many sequences, for some genes both parental copies are still carried in maize (Schnable *et al.*, 2009). In these cases, it is not always clear whether both genes are still under selection or if one has become a pseudogene. Within maize there are also many examples of presence absence variation and it is unclear how many such dispensable genes are functionally important (Hufford *et al.*, 2021; Schnable, 2020).

Multiple methods of identifying evolutionarily constrained regions have been developed. Many existing methods rely on aligning the genomes of multiple species and identifying sites which are conserved across them (Davydov *et al.*, 2010b; Siepel *et al.*, 2005). These approaches are limited by the ability to align other species to the reference species of interest, an especially difficult task for plant genomes. Among sequenced mammalian genomes, the median percentage syntenic between two species is greater than 75%, while for angiosperm species the comparable value is less than 12.5% (Zhao and Schranz, 2019). For maize only around 10% of the genome aligns to sorghum, one of its closest relatives. These approaches can only identify conservation on deep time scales and can not distinguish conservation only within one species. It is also not obvious how to treat regions of the genome that align multiply to other species, as occurs in maize due to being an ancestral allopolyploid. Another way to identify constrained regions is to use population genetic data. This has the advantage of being applicable in any region of the genome where polymorphisms can be called. In addition, population genetic data has the potential to identify conservation unique to one species. A population genetic method that uses a support vector machine (SVM) has previously been used to identify novel regions of constraint in the human genome (Schrider and Kern, 2015). We apply a similar method to maize population genetic data to identify novel regions of constraint in the maize genome.

In so doing, we address multiple questions about evolutionary genetics. First, do known or putative functional elements show evidence of selective constraint? Additionally, are there putative functional elements that appear to be unconstrained? Second, in the case of duplicated genes, can we identify pairs where one member is potentially pseudogenized and exhibits lower selective constraint? And third, can we use machine learning and population genetic data to identify novel constrained regions in the maize genome which are not detected by comparative methods?

3.2 Methods

3.2.1 Site frequency spectrum

We derived a site frequency spectrum (SFS) for 10kb windows in the maize genome from 772 maize inbred lines included in HapMap and the 282 association panel (Bukowski *et al.*, 2018; Flint-Garcia *et al.*, 2005). We filtered out samples that were different accessions of the same lines, teosinte lines, and lines that appeared to be identical based on relatedness. We aligned short read data from the 772 lines to v5 of the B73 maize reference genome using BWA-MEM (Hufford *et al.*, 2021; Li, 2013). We then used Sentieon Haplotyper to call SNPs with default settings (Freed *et al.*, 2017). We used GATK v3 to filter out sites with quality normalized by allele depth less than 2, Fisher strand bias greater than 60, mapping quality less than 40, mapping quality rank sum less than -12.5, and read position rank sum less than -8, and sites for individuals that were called heterozygous (DePristo *et al.*, 2011). In order to remove SNPs that may have been incorrectly called due to paralogy or that were within repetitive sequence we made a mask file using SNPable to identify uniquely mapping 35-mers with other parameters default and obtained the RepeatMasker mask for the B73 genome (Li, 2009). We used VCFtools version 0.1.14 to select only SNPs within the SNPable mask and outside of regions identified as repetitive by RepeatMasker and then get the counts of each allele (Danecek *et al.*, 2011). We downsampled the total count of each SNP to 458, the 10% quantile across SNPs. SNPs which were genotyped in fewer than 458 individuals were excluded from our analysis. The downsampled SNPs were used to obtain a site frequency spectrum (SFS) for each 10 kb window in the maize genome (Quinlan and Hall, 2010). As we do not know the ancestral state of each SNP, we recorded counts of the minor allele at each SNP to construct a "folded" SFS. We filtered out all windows where less than 10% of the window was within the SNPable mask.

3.2.2 Machine learning

We used a support vector machine (SVM) to identify regions of constraint in the genome. Using the bins of counts of SNPs at each frequency in the SFS as features, we trained the SVM to predict constrained and unconstrained windows across the genome. To identify constrained windows, we used genomic evolutionary rate profiling (GERP) scores for B73 (Davydov *et al.*, 2010b; Hufford *et al.*, 2021). GERP is a comparative method that identifies elements conserved across evolutionary history. We classified any window where GERP predicted that >25% of the window was constrained as a constrained window. To identify unconstrained windows we also used annotations of gene models and accessible chromatin from ATAC-seq (Hufford *et al.*, 2021). Windows that had no constrained GERP sites and did not overlap with genes or open chromatin were classified as unconstrained. The remaining ambiguous windows were not included in our training set.

Before running the algorithms, we separated 1/3 of the high confidence set into a training set, 1/3 into a validation set, and the remaining 1/3 into a test set. The training and validation sets were used while developing our implementation of the SVM while the test set was reserved to measure the performance of the final SVM.

We used the packages *kernlab* and *caret* in R to run the SVM (Karatzoglou *et al.*, 2004; Kuhn, 2008). The parameter sigma was selected based on the default heuristic in *kernlab* and the parameter C was chosen based on a search using powers of two from 2^{-2} to 2^{7}. We also set "caret" to compute class probabilities for whether a window was constrained or unconstrained. We assessed performance using a confusion matrix and an Receiver Operating Characteristic (ROC) curve estimated with with the R package “ROCR”.

3.2.3 Analyses

We compared the results of the machine learning approach to several annotations of the maize genome. In addition to the previously mentioned annotations of genes and open chromatin from ATAC-seq, we also used expression data from the NAM founders, a list of genes in B73 syntenic to *Sorghum bicolor* (Hufford *et al.*, 2021), and a list of putative enhancers (Oka *et al.*, 2017). As the list of enhancers was on the coordinates of v4 of the b73 genome, we used the software liftOver (Kuhn *et al.*, 2013) to port them to v5.

We calculated the percentage of genes, coding sequence, open chromatin, and enhancers that were covered by windows predicted to be constrained and the percentage of windows predicted to be constrained or unconstrained covered by annotations. We used loess regression to fit the relationship between a window overlapping an annotation and its predicted constraint with the package for genes and open chromatin. This was not possible computationally for enhancers, possibly because of the small number of windows with enhancers, so instead we used a generalized additive model.

To test whether genes in constrained windows had higher expression, used measures of expression from RNA-seq performed on seven tissues of B73 (Hufford *et al.*, 2021). We measured expression as reads per kilobase of transcript, per million mapped reads (RPKM). For genes with multiple transcripts, we used the transcript with the greatest RPKM. We fit a generalized additive model for all annotated genes in B73 with $mean\_constraint \sim log(RPKM + 1)$, using the mean constraint of all windows overlapping a gene, weighted by the length of the overlap. We also tested whether genes in constrained windows were more likely to be associated with a protein with measured expression. Using proteome data from (Walley *et al.*, 2016), we classified genes according to whether or not they had measured protein expression in any tissue or timepoint. We then used a chi-square test to ask whether genes in windows that were predicted to be constrained were equally likely to have an associated protein as genes in unconstrained windows.

For maize genes that were the result of a duplication since maize split with the common ancestor of maize and sorghum, we compared the predicted constraint of the two copies to determine if both copies were equally constrained or if one copy had experienced relaxed constraint. We identified gene pairs where the average probability of constraint was > 0.95

in overlapping windows for one gene and < 0.05 in the other as putative instances of pseudogenization. We then used a Wilcoxon paired test to ask whether putatively pseudogenized genes had fewer GWAS hits with 50 kb than their constrained counterparts using the results of a meta-analysis of multiple maize GWAS (Wallace *et al.*, 2014). As GWAS hits were originally on v2 of the maize genome, we used liftOver to port them to v5. For each pair of gene copies where one had a probability of constraint > 0.5 and the other had a probability of constraint < 0.5, we compared gene expression levels and asked if the copy with greater predicted constraint had greater expression. We filtered out genes with no recorded expression.

We also asked if genes present in B73 but but missing in one of the other founders (near-core genes) or more than one of the other founders (dispensable genes) were less likely to fall in windows predicted constrained than genes found in all founders (core genes) (Hufford *et al.*, 2021). As dispensable genes must by definition be under less constraint, we predicted they should have lower predicted constraint.

We looked at which windows were predicted unconstrained by the SVM and constrained by GERP and vice versa to find evidence of either recent changes in selection within maize or windows where one or the other method fails. We limited this analysis to a high confidence set of windows that had either a $> 95\%$ probability of being constrained or a $< 5\%$ probability of being constrained. We called a window constrained based on GERP if greater than 15% of it was covered by constrained GERP elements and unconstrained if it was covered by $<$ 1% GERP elements. This is different from the classifications we used in defining windows when training the SVM. We reduced the percentage of the window that must have positive GERP scores for the window to be called constrained from 25% to 15% because at the more strict threshold we only identified one such window. The windows called unconstrained by GERP included those with few or no GERP scores due to not aligning to other genomes. To consider only windows that appeared to be unconstrained based on GERP in regions where there were scores, we looked at only windows where at least half of the sites had GERP scores. We used a Mann-Whitney U test to ask whether windows that are called unconstrained by the SVM but constrained by GERP have fewer GWAS hits within 50 kb than all windows with a $> 50\%$ probability of being constained. Similarly, we also asked whether windows that are called constrained by the SVM but unconstrained by GERP have more GWAS hits within 50 kb than all windows with $< 50\%$ probability of being constrained.

3.3 Results

We called SNPs from 772 inbred lines and obtained the site frequency spectrum (SFS) for each 10 kb window in the maize genome. Using GERP and functional annotations we identified a subset of windows we could classify as constrained or unconstrained with high confidence, and used these windows to train an SVM to predict constraint based on the SFS.

3.3.1 SVM performance

We assessed the performance of the SVM with multiple metrics. When evaluated on the entire test set, the SVM has an overall accuracy of 89.74%, an accuracy of 89.58% when predicting unconstrained windows and an accuracy of 90.22% when predicting constrained windows (Fig. 3.1 A). When considering windows in the test set that the SVM classifies with probability greater than 95%, the SVM has an overall accuracy of 95.80%, an accuracy of 95.92% when predicting unconstrained windows and an accuracy of 95.46% when predicting constrained windows. (Fig. 3.1 B). The ROC curve shows that when keeping false positives below 5% the SVM can identify true positives at a rate greater than 80% (Fig. 3.1 C).

Much of the maize genome is repetitive and both lies outside the scope of this method as well as most likely not being under evolutionary constraint. At greater than 95% confidence the SVM predictions cover 17.07% of the total genome (Table 3.1). In comparison, 6.0% of the total genome is covered by GERP scores. Additionally, 39.40% of sites with GERP scores are covered by windows with greater than 95% confidence and 64.41% are covered by windows with greater than 80% confidence. When considering only the non-repetitive portion of the genome, 28.80% is covered by SVM predictions at greater than 95% probability (Table 3.2). When setting the cutoff at 80%, slightly more than half of the non-repetitive genome is covered.

3.3.1.1 Constraint in the maize genome

To measure how well the SVM predictions are capturing functional elements, we asked what fraction of various functional annotations is captured within predicted constrained windows. When looking at constrained windows across the genome, the distribution of the number of genic base pairs is shifted to the right compared with unconstrained windows (Fig. 3.2). Of base pairs in windows that passed our quality thresholds, 72.55% of genic base pairs lie within predicted constrained windows. The respective percentages for open chromatin and enhancers are 79.06% and 73.78%. Given that overall 41.19% of the genome is predicted to be constrained, all of these functional annotations have higher percentages of overlap with constrained window than the genome wide baseline.

Windows with non-zero overlap with genes had a higher median probability of being constrained than windows that did not overlap with genes (Fig. 3.3). Even in windows with only small number of genic base pairs, the median probability of constraint is more than 10% higher than in windows with zero genic base pairs. Of windows with more than 100 genic base pairs, the majority are predicted to be constrained. The results of a loess regression on windows with non-zero overlap with genes showed that windows with an intermediate number of genic base pairs tended to have a higher probability of constraint. This effect is largely driven by a rapid drop in the probability of constraint in windows that have close to 10,000 genic base pairs. The results of a loess regression with base pairs of coding sequence as the explanatory variable also showed that intermediate overlap with coding sequence (between approximately 2,500 and 5,000 base pairs) had the highest probability of being constrained. However, out of 32,777 windows that overlapped with coding sequence, only 62 included more than 5,000 base pairs of coding sequence. The probability of constraint increases with the number of base pairs of open chromatin. In windows overlapping with enhancers, windows with high overlap with enhancers were the most likely to be constrained, but there was a dip in the probability of constraint at intermediate overlap with enhancers.

Using a dataset of gene expression in multiple tissues in B73, we analyzed the relationship between predicted constraint and gene expression. We found that when looking at genes with expression in the dataset we used, genes with higher expression pooled across multiple tissues in B73 tended to be in windows with higher predicted constraint (Fig. 3.4). The Spearman's rank correlation between probability of constraint and log read count was significant ($rho = 0.190$, p-value $<$ 2.2e-16). We found that the difference in mean constraint between genes with zero expression and nonzero expression was much smaller than the difference between lowly expressed and highly expressed genes. When looking at protein expression data, a chi-square test showed that genes in constrained windows were more likely to be associated with a known expressed protein than genes in unconstrained windows, with 35.8% of genes in constrained windows being associated with a protein and 31.3% of genes in unconstrained windows being associated with a protein (p=0.0002732).

We compared members of gene pairs where one member was predicted constrained and the other was predicted constrained. When looking at all gene pairs where one had a predicted constraint > 0.5 and the other < 0.5 and there was measured expression for both genes, we found a significant correlation between the difference in probability of constraint and the difference in expression level (Pearson's correlation = 0.212 [0.083, 0.336], p-value = 0.001542. Spearman's correlation = 0.262, approximate p-value = 8.937e-05). If we narrow to a much more conservative set, we identified 18 gene pairs where one gene copy appears to have experienced pseudogenization and the other appears to still be constrained. A Wilcoxon paired test showed that the unconstrained member of each pair was within 50 kb of fewer GWAS hits than the constrained member.

When comparing core, near-core, and dispensable genes, we find that the percent overlap with predicted constrained windows is greatest for core genes, followed by near-core, private, and then dispensable genes (Table 3.3). Within windows that pass quality thresholds, 79.27% of core genes are in windows called constrained while that value is only 45.54% for dispensable genes. The median probability of constraint of windows overlapping with core genes is 91.6% and the median probability of constraint of dispensable genes is 40.6% (Table 3.4). Dispensable genes are also more likely to lie within windows removed by quality thresholds by the SVM (67.5%) than core genes (18.0%).

As the SVM relies only on population genetic data, it may be able to identify maize-specific constraint more accurately than a comparative genomic method such as GERP. Comparative genomic methods are limited to regions where the genome of the species of interest can align to other species. Additionally, comparative methods identify constraint on a deep evolutionary time scale that spans multiple species, not constraint within the more recent history of one species. When looking at windows called by the SVM with > 95% confidence, we identified 23 windows that were called unconstrained by the SVM and constrained by GERP and 3,852 windows that were called constrained by the SVM and unconstrained by GERP. Of the 213,190 windows called unconstrained by GERP where at least 50% of the window had GERP scores, we identified 126 windows called constrained by the SVM at greater than 95% confidence. To assess whether the SVM or GERP was correct in cases where they disagreed, we asked whether windows predicted constrained by the SVM and unconstrained by GERP were within 50 kb of more GWAS hits and vice versa. Based on Mann-Whitney U tests, windows only predicted unconstrained by the SVM and constrained by GERP are near fewer GWAS hits than all windows with > 50% probability of being constrained (median svm predicted unconstrained = 1, median all constrained = 2, p=0.0187) and windows predicted constrained by the SVM and unconstrained by GERP similarly are close to more GWAS hits than all unconstrained windows (median svm predicted constrained GERP predicted unconstrained = 2, median all SVM predicted unconstrained = 1, p < 2.2e-16).

Probability	Constrained	Unconstrained	Total
80%	11.14%	17.63%	28.76%
95%	5.84%	11.22%	17.07%

Table 3.1: Percentage of the genome covered by SVM predictions at two different levels of confidence.

Probability	Coverage
80%	50.31%
95%	28.80%

Table 3.2: Percentage of the non-repetitive genome covered by SVM predictions at two different levels of confidence.

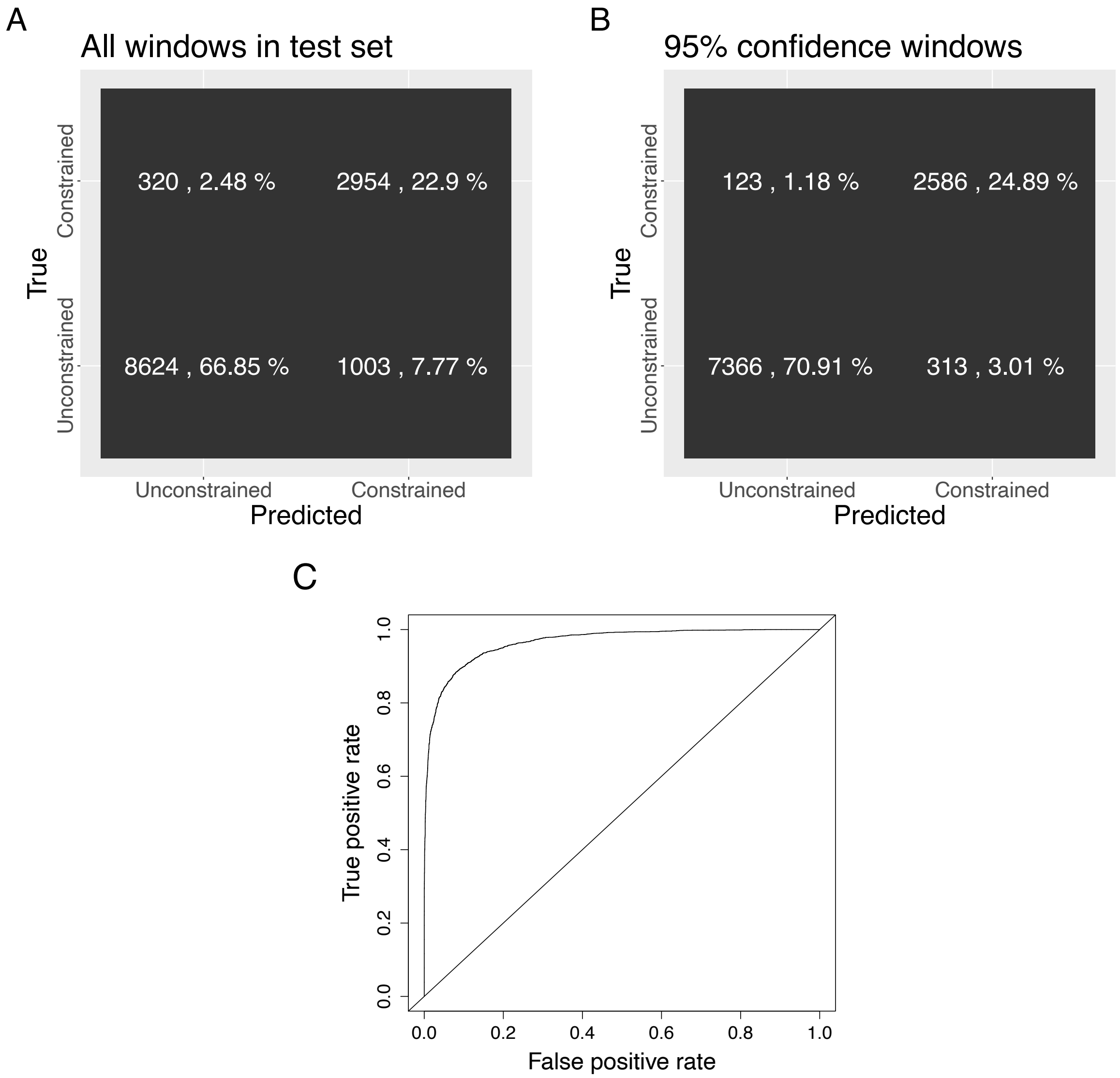

Figure 3.1: A) Confusion matrix of the SVM run on the test set. B) Confusion matrix of the SVM run on the test set only including high confidence (>0.95 probability) predictions. C) ROC curve for the SVM run on the test set.

Gene class	Constrained	Unconstrained	Unclassified	Sample size
Core gene	65.0%	17.0%	18.0%	38279
Near-core	39.5%	16.4%	44.1%	4823
Dispensable	14.8%	17.7%	67.5%	16216
Private	35.6%	33.3%	31.1%	444

Table 3.3: Percentage of core, near-core, dispensable, and private genes present in B73 that overlap with predicted constrained windows.

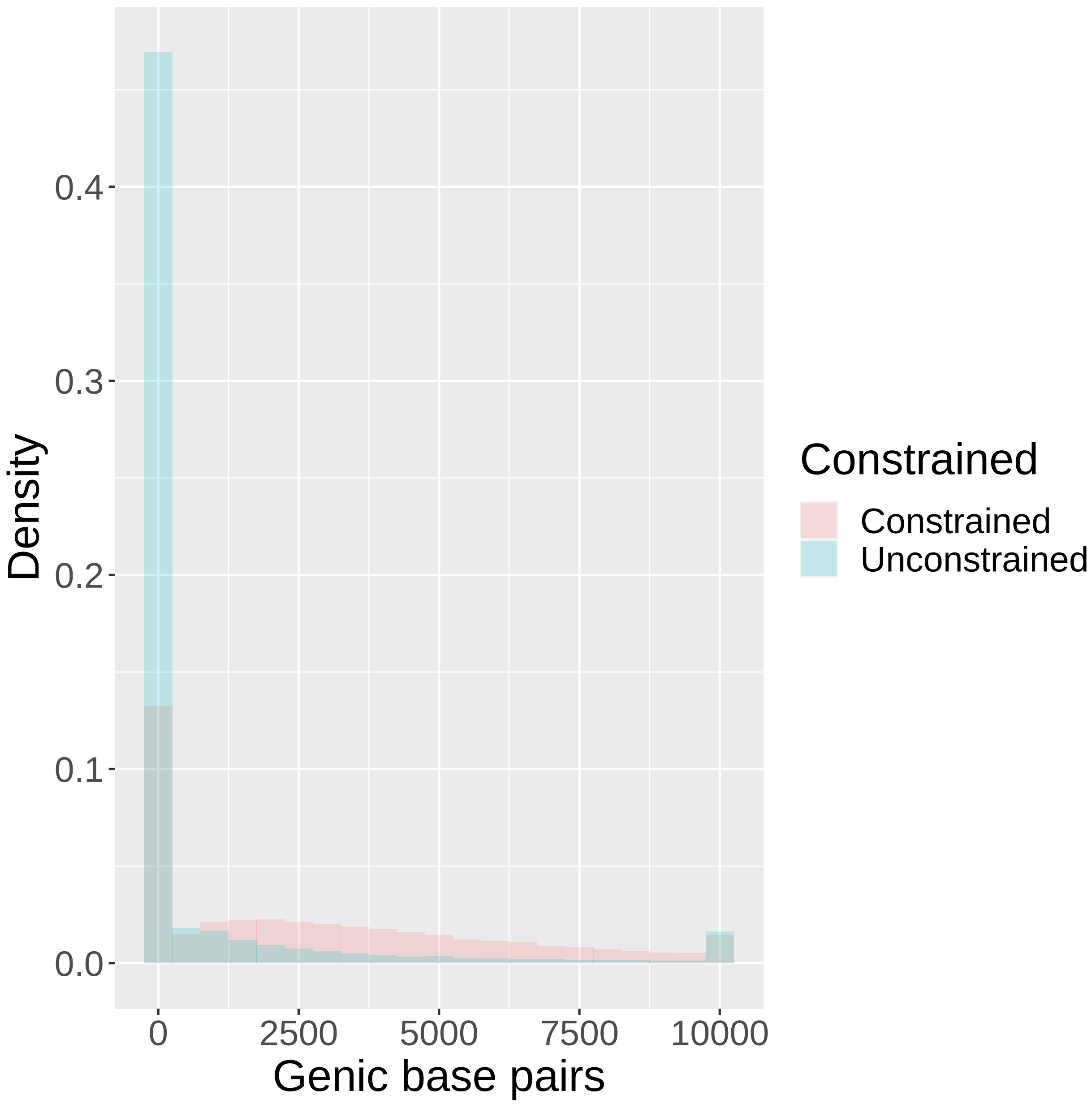

Figure 3.2: Histograms of the distribution of genic base pairs in predicted constrained and unconstrained windows.

Gene class	Median probability of constraint
Core gene	91.6%
Near-core	85.9%
Dispensable	40.6%
Private	53.1%

Table 3.4: Median probability of being constrained of windows that overlap core, near-core, dispensable, and private genes present in B73.

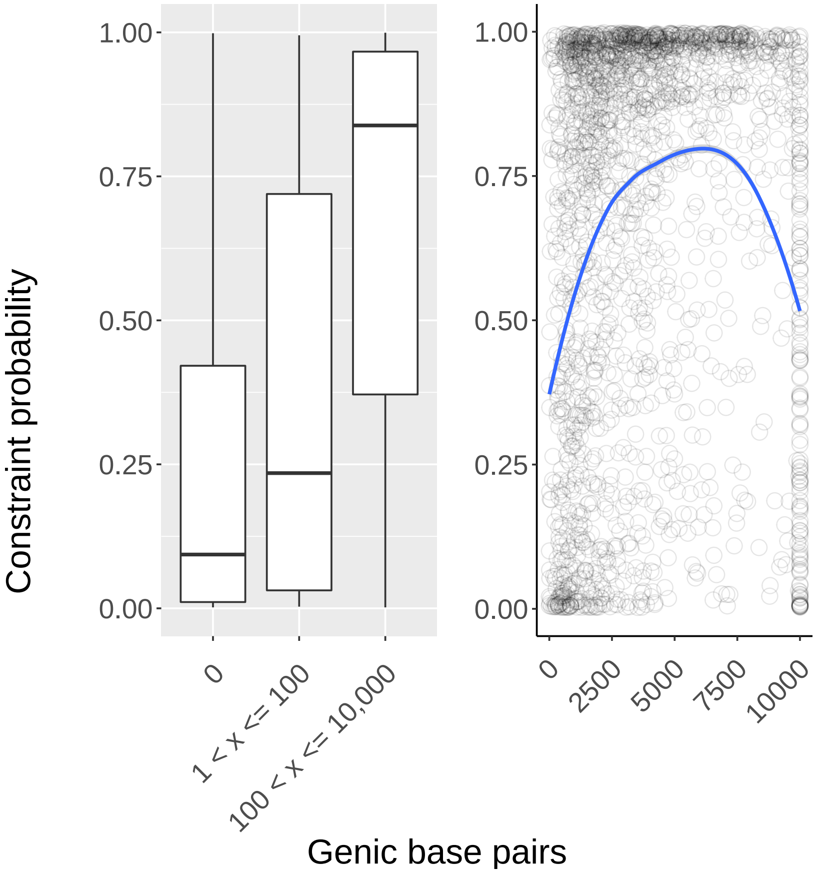

Figure 3.3: On the left are box plots of the probabilities of constraint for windows without genic base pairs, between 1 and 100 genic base pairs, and greater than 100 genic base pairs. On the right is the curve of a loess regression predicting probability of constraint based on the number of genic base pairs in a window. Circles represent a random sample of 2,000 windows.

3.4 Discussion

We show that an SVM using population genetic data can accurately predict evolutionary constraint in a plant species. In the test set the SVM has an accuracy of 89.74%, which is comparable to what was found when this method was applied to humans (Schrider and Kern, 2015). Compared to using comparative genomic methods that require outgroup alignment, the SVM substantially increases the fraction of the genome we can assess for constraint. While GERP scores allow estimation of constraint in only 6.0% of the maize genome, we can assess constraint for 28.76% of the genome using the SVM with 80% confidence. With much higher confidence (95% accuracy) we are still able to assay nearly three times as much of the genome (17.07%) than previously possible.

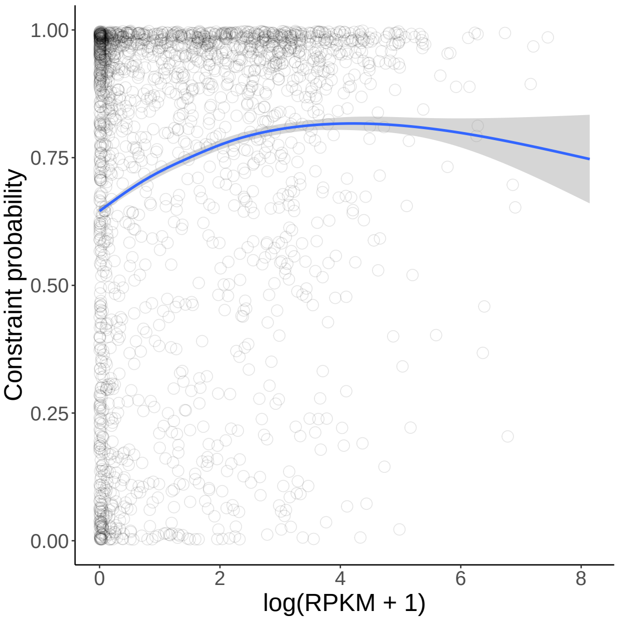

Figure 3.4: Results from a generalized additive model of probability of constraint on the log of RPKM + 1 for genes with expression greater than 0.

The accuracy of the SVM's predictions is further supported by the overlap between functional annotations and windows predicted constrained, which is greater than the baseline percentage across the genome for genes, open chromatin, and enhancers. Given this result, it is not surprising that a loess regression shows that the number of functional base pairs in windows overlapping these annotations is a good predictor of constraint (Fig. 3.3). What was more surprising was that the probability of constraint in windows with the highest number of genic base pairs was lower than in windows with intermediate numbers of genic base pairs. We can think of two possible explanations for this. First, windows that entirely overlap with genes may be more likely to be overlapping with introns. The median length of a gene in maize is slightly above 2.7 kb and fewer than 10% of maize genes are larger than 9.9 kb (Hufford *et al.*, 2021). Windows that overlap larger genes may be more likely to overlap with introns. Second, large genes may be more likely to contain unconstrained sequence, whether composed of introns or exons, or to be misannotated. The loess regression of probability of constraint on number of base pairs of coding sequence provides evidence to support these explanations. The vast majority of windows that overlap with coding sequence contain fewer than 5,000 base pairs of coding sequence, and when considering those windows the relationship between number of coding base pairs and probability of constraint appears to be essentially positive.

Population genetic methods have the potential to identify more recent selection than comparative genomic methods, which by design identify selection that has been consistent across multiple species. If a region of the genome has only recently become constrained within a species, or has recently become unconstrained, a comparative genomic method may incorrectly identify constraint. We identified windows where the SVM and GERP disagreed about constraint. To investigate which method was more likely to be correct, we compared the number of GWAS hits near windows where the two methods disagreed, hypothesizing that truly constrained windows would be closer to more functional base pairs, where mutations could contribute to phenotypic variance. We found that windows the SVM called constrained were closer to more GWAS hits than windows the SVM called unconstrained, even when the GERP classification disagreed. This supports the idea that the SVM is accurately predicting constraint within maize, possibly due to changes in constraint between other species and maize.

Higher gene expression was associated with higher predicted constraint among genes with measured expression. This result is in line with previous work — for example, core genes present in all maize lines have higher gene expression than dispensable genes (Hufford *et al.*, 2021). The correlation between predicted constraint and expression was significant but not very high, which is not surprising as we would not expect gene expression to be the main factor explaining constraint. If gene expression is under stabilizing selection, the optimum value for any given gene is presumably dependent on the relationship between expression and phenotype and would not always be higher for more constrained genes. For many genes in primates, mice, and flies this does appear to be the case, although evidence is lacking for plants (Romero *et al.*, 2012; Lemos *et al.*, 2005). More surprisingly, we found that genes with zero measured expression were not less likely to be predicted constrained by the SVM than genes with expression. It may be the case that many of the genes with no expression are actually expressed in a tissue, timepoint, or environment not captured in the data we used. Including these genes in the unexpressed category may thus be introducing noise into the data that prevents us from seeing any true signal. Interestingly, in Walley *et al.* (2016) the authors found that more than 95% of syntenic genes were expressed across 23 tissues. In the expression data we used from Hufford *et al.* (2021), the authors measured expression across seven tissues and only detected expression for 60.7% of syntenic genes, which supports the idea that some of the genes with zero expression in this data may actually be expressed in tissues not measured in their study.

Comparing the predicted constraint of windows containing core and non-core genes revealed that windows with core genes were more likely to be called constrained than windows with near-core or dispensable genes. This makes intuitive sense in that if a gene is not present in all members of a species it can not be essential and is not as constrained as a core gene might be. We also find that windows containing dispensable genes are more likely to not pass our thresholds for mappability. Windows with dispensable genes will have different numbers of genotypable sites in different individuals — in an individual where that gene is not present, no SNPs can be ascertained within the gene. This may affect the shape of the SFS in windows with dispensable genes relative to those with core genes. While there are fewer disposable genes we can predict constraint for, when only considering windows that passed these thresholds we still find that the median probability of constraint of windows including core genes is more than twice that of windows including dispensable genes. From these results we conclude that while there are some non-core genes the SVM method can not predict constraint for, overall dispensable genes are less evolutionarily constrained than core genes. Other work also suggests that dispensable genes are less likely to be functional, including that they are less likely to be expressed and that they overlap less than expected with genes in maize and *Arabidopsis thaliana* whose mutants are known to have phenotypic effects (Hufford *et al.*, 2021; Schnable, 2020; Bush *et al.*, 2014; Liang *et al.*, 2019).

We used the SVM to investigate constraint among gene pairs present in maize as a result of the whole genome duplication in the ancestor of maize. Among gene pairs with two copies in maize and only one in maize's relative sorghum, if one gene was predicted to be constrained and the other predicted to be unconstrained we found a significant positive correlation between the difference in the probability of constraint and expression of each gene. This provides evidence that the SVM is accurately discriminating which copy of a gene pair present in maize following duplication is more constrained. We identified a high confidence set of gene pairs where one gene copy was predicted to be constrained and the other unconstrained both with probability greater than 95% and provided further evidence for these predictions by showing that the gene copy predicted constrained tended to be close to more GWAS hits compared to the unconstrained copy. Following duplication, multiple outcomes are possible for duplicated genes, including pseudogenization (a relaxation of constraint), subfunctionalization, neofunctionalization, and the maintenance of both copies (all of which should lead to continued constraint) (Innan and Kondrashov, 2010). Using predictions from the SVM, we are able to provide evidence to assess which fate a duplicated gene pair has experienced.

These predictions may be useful for multiple types of research. First, for population genetic analyses, it is often necessary to identify regions under purifying selection or that are evolving neutrally. The SVM predictions could be used to select regions for use in population genetic analyses. Second, these results might be useful for prioritizing sites where mutations have phenotypic effects. We find that while most elements annotated as functional are within windows called constrained, some are not and we provide additional evidence that the SVM is accurately calling these windows unconstrained. Additionally, we find 4,934 windows predicted constrained that do not overlap with genes, open chromatin, or enhancers, some of which may include currently unknown regulatory sequence. This contributes to the body of research on which functionally annotated regions are in fact relevant to phenotype and under evolutionary constraint. Answering this question is important for multiple topics in biology, including development of genome annotations, evolution following genome duplication, and dispensable genes and the pan-genome. While the SVM provides probabilities of windows being constrained and we are able to support the accuracy of those probabilities over many windows using additional sources of data, these classifications do not exclusively rule out the possibility of the SVM being incorrect in individual windows. As accuracy is 95.90% in the test set, which likely includes some of the most constrained and unconstrained, and possibly easiest to predict, windows, some windows will invariably be called incorrectly. This can be mitigated by focusing on high confidence predictions. As noted in Schnable (2020) about gene annotations, validating each individual annotation for function would require testing mutants in each annotation in the many environments an individual could realistically encounter. Instead, combining multiple lines of evidence may be the most helpful strategy for assessing biological function within categories or in individual regions of the genome.

In conclusion, we show that an SVM can accurately predict evolutionary constraint in maize using population genetic data, that these results cover more of the genome than comparative genomic methods for the same purpose, and that these predictions are useful for addressing evolutionary questions.

3.5 Acknowledgements

A.I.H. was supported by National Science Foundation Graduate Research Fellowship 1650042 and the UC Davis Dept. of Plant Sciences.

References

Aguirre, J. D., E. Hine, K. McGuigan, and M. W. Blows, 2014 Comparing g: multivariate analysis of genetic variation in multiple populations. Heredity **112**: 21–29.

Allard, R. W. and A. D. Bradshaw, 1964 Implications of genotype-environmental interactions in applied plant breeding. Crop Science **4**: 503–508.

Barrett, R. D. H., S. M. Rogers, and D. Schluter, 2008 Natural selection on a major armor gene in threespine stickleback. Science **322**: 255–257.

Bradshaw, A., 1965 Evolutionary Significance of Phenotypic Plasticity in Plants. In *Advances in Genetics*, volume 13, pp. 115–155, Elsevier.

Brunner, S., K. Fengler, M. Morgante, S. Tingey, and A. Rafalski, 2005 Evolution of DNA Sequence Nonhomologies among Maize Inbreds. The Plant Cell **17**: 343–360.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown, *et al.*, 2009 The genetic architecture of maize flowering time. Science **325**: 714–718.

Bukowski, R., X. Guo, Y. Lu, C. Zou, B. He, *et al.*, 2018 Construction of the third-generation Zea mays haplotype map. GigaScience **7**.

Bush, S. J., A. Castillo-Morales, J. M. Tovar-Corona, L. Chen, P. X. Kover, *et al.*, 2014 Presence-Absence Variation in A. thaliana Is Primarily Associated with Genomic Signatures Consistent with Relaxed Selective Constraints. Molecular Biology and Evolution **31**: 59–69.

Bürkner, P.-C., 2017 brms: An r package for bayesian multilevel models using stan. Journal of Statistical Software **80**: 1–28.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature Genetics **22**: 231–238.

Covarrubias-Pazaran, G., 2016 Genome assisted prediction of quantitative traits using the r package sommer. PLoS ONE **11**: 1–15.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, *et al.*, 2011 The variant call format and VCFtools. Bioinformatics **27**: 2156–2158.

Davydov, E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, *et al.*, 2010a Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Computational Biology **6**.

Davydov, E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, *et al.*, 2010b Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Computational Biology **6**: e1001025.

Dell'Acqua, M., D. M. Gatti, G. Pea, F. Cattonaro, F. Coppens, *et al.*, 2015 Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in zea mays. Genome Biology **16**: 167.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics **43**: 491–498.

Des Marais, D. L., K. M. Hernandez, and T. E. Juenger, 2013 Genotype-by-environment interaction and plasticity: Exploring genomic responses of plants to the abiotic environment. Annual Review of Ecology, Evolution, and Systematics **44**: 5–29.

El-Soda, M., W. Kruijer, M. Malosetti, M. Koornneef, and M. G. M. Aarts, 2015 Quantitative trait loci and candidate genes underlying genotype by environment interaction in the response of *A rabidopsis thaliana* to drought: Genetics of drought response in arabidopsis. Plant, Cell & Environment **38**: 585–599.

Fan, L., L. Quan, X. Leng, X. Guo, W. Hu, *et al.*, 2008 Molecular evidence for post-domestication selection in the Waxy gene of Chinese waxy maize. Molecular Breeding **22**: 329–338.

Feuk, L., A. R. Carson, and S. W. Scherer, 2006 Structural variation in the human genome **7**: 85–97.

Filiault, D. L. and J. N. Maloof, 2012 A genome-wide association study identifies variants underlying the arabidopsis thaliana shade avoidance response. PLoS Genetics **8**: e1002589.

Finlay, K. and G. Wilkinson, 1963 The analysis of adaptation in a plant-breeding programme. Australian Journal of Agricultural Research **14**: 742.

Flint-Garcia, S. A., A.-C. Thuillet, J. Yu, G. Pressoir, S. M. Romero, *et al.*, 2005 Maize association population: a high-resolution platform for quantitative trait locus dissection: High-resolution maize association population. The Plant Journal **44**: 1054–1064.

Freed, D., R. Aldana, J. A. Weber, and J. S. Edwards, 2017 The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. preprint, Bioinformatics.

Fuentes, R. R., D. Chebotarov, J. Duitama, S. Smith, J. F. De la Hoz, *et al.*, 2019 Structural variants in 3000 rice genomes. Genome Research **29**: 870–880.

Gage, J. L., D. Jarquin, C. Romay, A. Lorenz, E. S. Buckler, *et al.*, 2017 The effect of artificial selection on phenotypic plasticity in maize. Nature Communications **8**: 1348.

Gates, D. J., D. Runcie, G. M. Janzen, A. R. Navarro, M. Willcox, *et al.*, 2019 Single-gene resolution of locally adaptive genetic variation in mexican maize.

Gaut, B. S. and J. F. Doebley, 1997 DNA sequence evidence for the segmental allotetraploid origin of maize. Proceedings of the National Academy of Sciences **94**: 6809–6814.

Ghavi-Helm, Y., A. Jankowski, S. Meiers, R. R. Viales, J. O. Korbel, *et al.*, 2019 Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nature Genetics **51**: 1272–1282.

Graur, D., Y. Zheng, N. Price, R. B. R. Azevedo, R. A. Zufall, *et al.*, 2013 On the Immortality of Television Sets: ”Function” in the Human Genome According to the Evolution-Free Gospel of ENCODE. Genome Biology and Evolution **5**: 578–590.

Hake, S. and J. Ross-Ibarra, 2015 Genetic, evolutionary and plant breeding insights from the domestication of maize. eLife **4**: e05861.

Houle, D., 1992 Comparing evolvability and variability of quantitative traits. Genetics **130**: 195–204.

Huang, K. and L. H. Rieseberg, 2020 Frequency, Origins, and Evolutionary Role of Chromosomal Inversions in Plants. Frontiers in Plant Science **11**: 296.

Hubisz, M., K. Pollard, and A. Siepel, 2018 *rphast: Interface to 'PHAST' Software for Comparative Genomics*. R package version 1.6.9.

Hudson, A. I., S. G. Odell, P. Dubreuil, M.-H. Tixier, S. Praud, *et al.*, 2022 Analysis of genotype-by-environment interactions in a maize mapping population. G3 Genes—Genomes—Genetics **12**, jkac013.

Hufford, M. B., A. S. Seetharam, M. R. Woodhouse, K. M. Chougule, S. Ou, *et al.*, 2021 De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science **373**: 655–662.

Innan, H. and F. Kondrashov, 2010 The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics **11**: 97–108.

Joseph, S. B. and D. W. Hall, 2004 Spontaneous Mutations in Diploid Saccharomyces cerevisiae. Genetics **168**: 1817–1825.

Josephs, E. B., 2018 Determining the evolutionary forces shaping *G* x *E*. New Phytologist **219**: 31–36.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis, 2004 **kernlab** - An *S4* Package for Kernel Methods in *R*. Journal of Statistical Software **11**.

Kawecki, T. J. and D. Ebert, 2004 Conceptual issues in local adaptation. Ecology Letters **7**: 1225–1241.

Kazan, K. and R. Lyons, 2016 The link between flowering time and stress tolerance. Journal of Experimental Botany **67**: 47–60.

Keightley, P. D. and M. Lynch, 2003 TOWARD A REALISTIC MODEL OF MUTATIONS AFFECTING FITNESS. Evolution **57**: 683–685.

Kiełbasa, S. M., R. Wan, K. Sato, P. Horton, and M. C. Frith, 2011 Adaptive seeds tame genomic sequence comparison. Genome Research **21**: 487–493.

Kirkpatrick, M., 2010 How and Why Chromosome Inversions Evolve. PLoS Biology **8**: e1000501.

Korte, A. and A. Farlow, 2013 The advantages and limitations of trait analysis with GWAS: a review. Plant Methods **9**: 29.

Kuhn, M., 2008 Building Predictive Models in *R* Using the **caret** Package. Journal of Statistical Software **28**.

Kuhn, R. M., D. Haussler, and W. J. Kent, 2013 The UCSC genome browser and associated tools. Briefings in Bioinformatics **14**: 144–161.

Lande, R., 1979 Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution **33**: 402–416.

Lemos, B., C. D. Meiklejohn, M. Cáceres, and D. L. Hartl, 2005 RATES OF DIVERGENCE IN GENE EXPRESSION PROFILES OF PRIMATES, MICE, AND FLIES: STABILIZING SELECTION AND VARIABILITY AMONG FUNCTIONAL CATEGORIES. Evolution **59**: 126–137.

Leushkin, E. V., G. A. Bazykin, and A. S. Kondrashov, 2013 Strong Mutational Bias Toward Deletions in the Drosophila melanogaster Genome Is Compensated by Selection. Genome Biology and Evolution **5**: 514–524.

Li, H., 2009 Snpable regions.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with bwa-mem.

Liang, Z., Y. Qiu, and J. C. Schnable, 2019 Distinct characteristics of genes associated with phenome-wide variation in maize (*Zea mays*). preprint, Bioinformatics.

Lieber, M. R., 2010 The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annual Review of Biochemistry **79**: 181–211.

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, *et al.*, 2011 FaST linear mixed models for genome-wide association studies. Nature Methods **8**: 833–835.

Mahmoud, M., N. Gobet, D. I. Cruz-DÃ¡valos, N. Mounier, C. Dessimoz, *et al.*, 2019 Structural variant calling: the long and the short of it **20**: 246.

Manzaneda, A. J., P. J. Rey, J. T. Anderson, E. Raskin, C. Weiss-Lehman, *et al.*, 2015 Natural variation, differentiation, and genetic trade-offs of ecophysiological traits in response to water limitation in *Brachypodium distachyon* and its descendent allotetraploid *B. hybridum* (Poaceae). Evolution **69**: 2689–2704.

Odell, S. G., A. I. Hudson, S. Praud, P. Dubreuil, M.-H. Tixier, *et al.*, 2022 Modeling allelic diversity of multi-parent mapping populations affects detection of quantitative trait loci. G3 .

Oka, R., J. Zicola, B. Weber, S. N. Anderson, C. Hodgman, *et al.*, 2017 Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. Genome Biology **18**: 137.

Okagaki, R. J., M. G. Neuffer, and S. R. Wessler, 1991 A deletion common to two independently derived waxy mutations of maize. Genetics **128**: 425–431.

Paaby, A. B. and M. V. Rockman, 2014 Cryptic genetic variation: evolution's hidden substrate. Nature Reviews Genetics **15**: 247–258.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, *et al.*, 2007 PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics **81**: 559–575.

Pyhäjärvi, T., M. B. Hufford, S. Mezmouk, and J. Ross-Ibarra, 2013 Complex patterns of local adaptation in teosinte **5**: 1594–1609.

Quinlan, A. R. and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**: 841–842.

R Core Team, 2020 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rodríguez-Álvarez, M. X., M. P. Boer, F. A. van Eeuwijk, and P. H. Eilers, 2017 Correcting for spatial heterogeneity in plant breeding experiments with p-splines. Spatial Statistics **23**: 52 – 71.

Rogers, A. R., J. C. Dunne, C. Romay, M. Bohn, E. S. Buckler, *et al.*, 2021 The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. G3 Genes|Genomes|Genetics **11**: jkaa050.

Romero, I. G., I. Ruvinsky, and Y. Gilad, 2012 Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics **13**: 505–516.

Romero Navarro, J. A., M. Willcox, J. BurgueÃ±o, C. Romay, K. Swarts, *et al.*, 2017 A study of allelic diversity underlying flowering-time adaptation in maize landraces **49**: 476–480.

Runcie, D. E. and L. Crawford, 2019 Fast and flexible linear mixed models for genome-wide genetics. PLOS Genetics **15**: e1007978.

Schnable, J. C., 2020 Genes and gene models, an important distinction. New Phytologist **228**: 50–55.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, *et al.*, 2009 The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science **326**: 1112–1115.

Schrider, D. R. and A. D. Kern, 2015 Inferring Selective Constraint from Population Genomic Data Suggests Recent Regulatory Turnover in the Human Brain. Genome Biology and Evolution **7**: 3511–3528.

Sgrò, C. M. and A. A. Hoffmann, 2004 Genetic correlations, tradeoffs and environmental variation. Heredity **93**: 241–248.

Sherrard, M. E., H. Maherali, and R. G. Latta, 2009 Water stress alters the genetic architecture of functional traits associated with drought adaptation in *Avena barbata*. Evolution **63**: 702–715.

Shomura, A., T. Izawa, K. Ebana, T. Ebitani, H. Kanegae, *et al.*, 2008 Deletion in a gene associated with grain size increased yields during rice domestication. Nature Genetics **40**: 1023–1028.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, *et al.*, 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research **15**: 1034–1050.

Smith, S. A. and J. W. Brown, 2018 Constructing a broadly inclusive seed plant phylogeny. American Journal of Botany **105**: 302–314.

Soreng, R. J., P. M. Peterson, K. Romaschenko, G. Davidse, J. K. Teisher, *et al.*, 2017 A worldwide phylogenetic classification of the Poaceae (Gramineae) II: An update and a comparison of two 2015 classifications: Phylogenetic classification of the grasses II. Journal of Systematics and Evolution **55**: 259–290.

Su, Z., A. Bernardo, B. Tian, H. Chen, S. Wang, *et al.*, 2019 A deletion mutation in TaHRC confers Fhb1 resistance to Fusarium head blight in wheat. Nature Genetics **51**: 1099–1105.

The ENCODE Project Consortium, 2012 An integrated encyclopedia of DNA elements in the human genome. Nature **489**: 57–74.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions **91**: 10.

Wallace, J. G., P. J. Bradbury, N. Zhang, Y. Gibon, M. Stitt, *et al.*, 2014 Association Mapping across Numerous Traits Reveals Patterns of Functional Variation in Maize. PLoS Genetics **10**: e1004845.

Walley, J. W., R. C. Sartor, Z. Shen, R. J. Schmitz, K. J. Wu, *et al.*, 2016 Integration of omic networks in a developmental atlas of maize. Science **353**: 814–818.

Wood, C. W. and E. D. Brodie, 2015 Environmental effects on the structure of the g-matrix. Evolution **69**: 2927–2940.

Xu, K., X. Xu, T. Fukao, P. Canlas, R. Maghirang-Rodriguez, *et al.*, 2006 Sub1a is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature **442**: 705–708.

Yang, N., J. Liu, Q. Gao, S. Gui, L. Chen, *et al.*, 2019 Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nature Genetics **51**: 1052–1059.

Yang, Q., Z. Li, W. Li, L. Ku, C. Wang, *et al.*, 2013 CACTA-like transposable element in *ZmCCT* attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proceedings of the National Academy of Sciences **110**: 16969–16974.

Yang, S., L. Wang, J. Huang, X. Zhang, Y. Yuan, *et al.*, 2015 Parentâprogeny sequencing indicates higher mutation rates in heterozygotes. Nature **523**: 463–467.

Zancolli, G., J. J. Calvete, M. D. Cardwell, H. W. Greene, W. K. Hayes, *et al.*, 2019 When one phenotype is not enough: divergent evolutionary trajectories govern venom variation in a widespread rattlesnake species p. 10.

Zhao, T. and M. E. Schranz, 2019 Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. Proceedings of the National Academy of Sciences **116**: 2165–2174.