A connected set of genes associated with programmed cell death implicated in controlling the hypersensitive response in maize


*Dept. of Plant Pathology, NC State University, Raleigh NC 27695-7616, USA.
§Botany and Plant Pathology, Purdue University, Lilly Hall, West Lafayette, IN 47907-2054, USA.
†Department of Computer Science, University of Missouri, Columbia MO 65211, USA.
‡USDA-ARS Plant Science Research Unit, Raleigh NC 27695, USA.
**Department of Plant and Soil Sciences, University of Delaware, Newark, DE, USA 19716, USA.
§§Dept of Crop Science, NC State University, Raleigh NC 27695-7620, USA.
††Informatics Institute, University of Missouri, Columbia, MO 65211, USA.

Author for correspondence

Peter Balint-Kurti
2572 Thomas Hall
Dept. of Plant Pathology,
NC State University
Raleigh NC 27695-7616
Tel 919 515 3516
Fax 919 856 4816
Peter.Balint-Kurti@ARS.USDA.GOV

1 These authors contributed equally to the work

Copyright 2012.
Abstract

Rp1-D21 is a maize auto-active resistance gene conferring a spontaneous hypersensitive defense response (HR) of variable severity depending on genetic background. We report an association mapping strategy based on the Mutant Assisted Gene Identification and Characterization (MAGIC) approach to identify naturally-occurring allelic variants associated with phenotypic variation in HR. Each member of a collection of 231 diverse inbred lines of maize constituting a high-resolution association mapping panel were crossed to a parental stock heterozygous for Rp1-D21 and the segregating F1 generation testcrosses were evaluated for phenotypes associated with lesion severity for two years at two locations. A genome-wide scan for associations with HR was conducted with 47,445 SNPs using a linear mixed model that controlled for spurious associations due to population structure. Since the ability to identify candidate genes and the resolution of association mapping are highly influenced by linkage disequilibrium (LD), we examined the extent of genome-wide LD. On average marker pairs separated by more than 10 kbp had an $r^2$ value of less than 0.1. Genomic regions surrounding SNPs significantly associated with HR traits were locally saturated with additional SNP markers in order to establish local LD structure and precisely identify candidate genes. Six significantly associated SNPs at five loci were detected. At each locus, the associated SNP was located within or immediately adjacent to candidate causative genes predicted to play significant roles in the control of programmed cell death and especially in ubiquitin pathway-related genes.

Keywords: Association analysis, Linkage disequilibrium, Maize, Hypersensitive response, quantitative trait, disease resistance
Introduction

The hypersensitive response (HR) mechanism is a widespread and important plant defense response. Characterized by a rapid, localized cell death around the point of attempted pathogen penetration, it is a form of programmed cell death (PCD) and is usually associated with an acute local resistance response and up-regulation of defense response pathways (Coll et al. 2011). HR and associated events are generally initiated by the products of “R” (resistance) genes, which trigger HR upon the recognition of specific pathogen-derived molecules or molecular events (Bent and Mackey 2007). The HR and related responses are generally associated with resistance to biotrophic rather than necrotrophic pathogens. Among the multiple classes of R-genes, those that encode proteins possessing a nucleotide-binding site (NBS) and leucine-rich repeat (LRR) are the predominant class (Bent and Mackey 2007).

The Rp1 locus on maize chromosome 10 carries multiple tandemly-repeated NBS-LRR paralogs, some of which confer resistance to specific races of maize common rust conferred by the fungus Puccini sorghi (Hulbert 1997) The locus is meiotically unstable due to a high frequency of unequal crossovers between paralogs (Sudupak et al. 1993). In one such case, unequal crossing over followed by intragenic recombination resulted in the formation of the chimeric gene Rp1-D21 (Collins et al. 1999; Smith et al. 2010). In the resulting gene product, the recognition and elicitation functions are partially uncoupled, causing the spontaneous formation of HR lesions on the leaves and stalks of the plant in the absence of pathogens. Rp1-D21 exhibits its lesion phenotype in a partially dominant and developmentally-dependent manner (Hu 1996; Smith et al. 2010). The strength of the phenotype is dependent on, among other things, genetic background (Chaikam et al. 2011; Chintamanani et al. 2010).
The *Rp1-D21* lesion phenotype can be used as a reporter for the identification of loci affecting the strength of HR triggered by *Rp1-D21*. Since the *Rp1-D21* lesion phenotype is an exaggerated defense response (Chintamanani *et al.* 2010), it is likely that many or all of these loci are also associated with variation in the wild type defense response. In previous work, (Chaikam *et al.* 2011; Chintamanani *et al.* 2010) a maize inbred line (H95) into which *Rp1-D21* was introgressed and maintained in a heterozygous condition (designated *Rp1-D21*-H95) was crossed with sets of lines from various mapping populations. By phenotyping the resulting F₁ progenies, several quantitative trait loci (QTL) modulating the HR conferred by *Rp1-D21* were identified. This approach, in which a mutant phenotype is used as a reporter to reveal previously undetectable genetically-controlled variation, has been termed Mutant-Assisted Gene Identification and Characterization (MAGIC; Johal *et al.* 2008). A similar approach was used to identify the *slm1* locus, a strong modulator of the *les23* lesion mimic gene in maize (Penning *et al.* 2004).

In conventional maize QTL studies using a structured population derived from a bi-parental cross of inbred lines, a maximum of two alleles are sampled; consequently, many loci important for controlling the trait of interest do not segregate in the mapping population and cannot be detected. This problem can be partially addressed by conducting multiple QTL analyses using populations derived from different bi-parental crosses, such as the maize NAM population (McMullen *et al.* 2009) or by using RILs derived from intermating multiple diverse lines or accessions (Cavanagh *et al.* 2008).

Alternatively, association mapping uses a population of diverse lines in which a wide genetic diversity is sampled. Just as with conventional QTL mapping, association mapping identifies QTL by seeking associations between the presence or absence of specific alleles and
variation in the trait of interest (Yu and Buckler 2006). Association mapping can not only assess a higher diversity of alleles, it also can lead to much more precise positional estimates due to the high number of recombination events accumulated during the historical diversification of the lines included in the population. An obstacle to genome-wide association mapping in low linkage disequilibrium (LD) populations has been the large number of markers required to detect marker-trait associations. Until recently this limited the search space to pre-determined candidate genes (Remington and Purugganan 2003). Advances in genomic technology have made it now possible to conduct genome-wide association studies (GWAS) in low LD populations.

Several maize association mapping populations have been constructed containing various sets of diverse lines (Liu et al. 2011; Lu et al. 2010; Yan et al. 2011; Yu et al. 2011). The most widely used of these consists of 302 inbred lines representing the diversity present in public-sector breeding populations around the world (Flint-Garcia et al. 2005). Here we will refer to this population as the “maize association population” Subsets of this population have been used for association mapping of several traits, including, maysin and chlorogenic acid accumulation (Szalma et al. 2005), flowering time (Thornsberry et al. 2001), kernel composition (Wilson et al. 2004) and flux in carotenoid biosynthesis pathways (Harjes et al. 2008). In all of these examples, a candidate gene approach was used in which genes already suspected of being involved in natural variation for the traits of interest were sequenced from each member of the population. Recently, 47,445 single nucleotide polymorphism (SNP) markers were scored on 279 of the 302 lines, enabling GWAS using this population (Cook et al. 2011; Ganal et al. 2011).

In this study, we combined the MAGIC and GWAS approaches to identify loci and genes associated with modulating the maize HR defense response. The $R_{p1}$-$D_{21}$-$H_{95}$ line, which is heterozygous for the $R_{p1}$-$D_{21}$ gene, was crossed to a subset (231 lines) of the maize association...
population, and the resulting $F_1$ families were evaluated in multiple environments. GWAS led to the identification of six SNP loci significantly associated with variation in the $Rp1-D21$ lesion phenotype. Since two of these SNPs were in high LD, this suggested that the effects of five causative genes were being detected. In each of the five cases, associated SNPs were localized within or adjacent to genes previously implicated in the control of programmed cell death and especially in the ubiquitin pathway associated with protein degradation. We also report on genome-wide LD decay in this association population as well as the extent of local LD decay around the significantly associated SNPs. This approach, combining MAGIC with GWAS offers great promise for the identification of alleles and loci associated with a variety of quantitative traits.
Materials and Methods

Plant materials

The $Rp1-D21$-H95 mutant line was created by crossing a $Rp1-D21$ variant and the maize inbred line H95; the F$_1$ was subsequently backcrossed to the H95 parent four times, while selecting for plants that formed spontaneous HR-like lesions. The $Rp1-D21$-H95 stock is maintained in a heterozygous state since $Rp1-D21$ homozygous plants are sterile.

The 302-line association population of maize is comprised of diverse inbred lines sampled from public-sector corn-breeding programs. Their pedigrees have been described elsewhere (Gerdes and Tracy 1993 and http://www.ars-grin.gov/). The $Rp1-D21$-H95 stock was crossed as a male to each of 231 lines (a subset of the 302 lines; Table S1, S2) to create a set of F$_1$ families, each of which segregated 1:1 for the presence/absence of $Rp1-D21$ but which were otherwise isogenic within a family. The selection of the 231 lines to use from the original 302 was based on the availability of genotypic data and sufficient testcross seed for phenotypic evaluation.

Field trials

Each of the 231 F$_1$ families was evaluated in four environments: in Clayton, NC and West Lafayette, IN, in the years 2009 and 2010. A randomized complete block design with two replicates in each location was used. Two rows of a constant genotype were planted around the edges of the field to eliminate border row effect. Standard fertilizer, pesticide and herbicide regimes were applied during the trial to ensure normal plant growth. Thinning to desired plant density and overhead irrigation were applied as required. At Clayton, NC, 10 kernels of each line
were sown in 2 m rows with an inter-row spacing of 0.97 m and a 0.6 m alley at the end of each plot, while at West Lafayette, IN, 18 seeds were sown in 6 m rows with an inter-row spacing of 0.76 m.

Phenotypic scoring

Each F₁ family segregated 1:1 for the presence/absence of Rp1-D21 but was otherwise isogenic. Within a family it was immediately obvious, by the presence or absence of lesions and the growth habit of the plant, which plants carried Rp1-D21 and which were wild-type (Figure S1). Fifteen lesion-associated traits were scored on each plot. For some of these traits only plants carrying Rp1-D21 were scored, while for others, both wild-type and mutant plants were measured and the mutant/wild-type ratio was calculated (see below). A description of each of the traits that were scored follows:

Traits derived from field observations:

**HR Lesion severity (LES):** LES was only measured on mutant plants. At both locations lesion severity scores were assigned based on a 0 to 10 scale, with 0 = no lesion and 10 = completely dead plant (Chaikam et al. 2011). Experiments were scored five times at West Lafayette, IN and six times at Clayton, NC starting one month after planting and continuing at approximately 10-14 day intervals.

We scored an aberrant defense response rather than disease in this case, but since the phenotypes observed are generally similar we used a widely-accepted statistic in plant pathology; standardized area under disease progress curve (sAUDPC), to measure quantitative levels of HR
The sAUDPC for LES was calculated for each environment as follows: The average value of two consecutive ratings was computed and multiplied by the number of days between the ratings. Values were summed over all intervals, and then divided by the total number of days over which evaluations were performed to determine the weighted average.

**Mutant to wild type height ratio (HTR):** Plant height data were collected from three representative mutant F_1 individuals and from three representative wild type F_1 individuals within each F_1 family. Height means were calculated for each class within each family and the HTR was calculated by dividing the average mutant type height to the average wild type height.

**Mutant to wild type stalk width ratio (SWR):** Stalk width immediately above the ear was measured from three representative mutant F_1 individuals and from three representative wild type F_1 individuals within each F_1 family. SWR was then calculated by dividing the average mutant type stalk width to the wild type average stalk width.

**Traits derived from image analysis**

At both the 3^{rd}/4^{th} and 7^{th}/8^{th} leaf stage, photographs were taken of two leaves per row for each row in each experiment, with the exception of the second replicate in Clayton 2009 which was not photographed. Images were taken using a Canon Rebel Xsi camera with a Gretag Macbeth Mini Color Checker included in the field of view. Images were preprocessed with custom algorithms written in C/C++ using the OpenCV library that: (1) standardize images by performing color correction, (2) identify leaves in the image, and (3) highlight necrotic leaves using spectral characteristics (Green et al. 2012). From this segmentation, a number of aggregate traits were computed:
**Percent necrotic lesions** (PCTLES): The percent necrotic lesions represented the proportion of the entire leaf identified as necrotic.

**Number of lesions** (NULES): The number of necrotic lesions trait is the count of the number of individual lesions highlighted in each image.

**Average necrotic lesion size** (LESSIZ): For average necrotic lesion size, the area of each detected lesion was measured in pixels with the average area computed and reported.

For each of these traits, averages for the 3/4th leaf and 7/8th leaf stages were obtained for each plot, and an average value across stages was calculated. A suffix of 4, 8 or AV was appended to the trait designation to indicate the stage to which it refers (e.g., LESSIZ4 LESSIZ8, LESSIZAV).

---

**Genotypic data**

We used genotype data from the Illumina maize 50k array, a set of 57,838 SNPs designed by Ganal et al (2011). Only the 47,445 SNP markers that mapped to defined single locations in the maize genome and had less than 20% missing data were used in the association analysis. Additional SNP markers developed by Dr. Ed Buckler’s research group (USDA-ARS, Cornell University) by a genotyping-by-sequencing (GBS) method (Elshire et al. 2011) were retrieved from http://www.panzea.org/dynamic/derivative_data/genotypes/Maize282_GBS_genos_imputed_20110.zip. GBS markers were analyzed for approximately 2 Mbp windows around SNPs from the 50k Illumina array data set that were detected as having significant associations with phenotypic traits measured in this study.
Statistical Analyses

Supplemental Files: Supplemental files 1-10 contain most of the phenotypic and genotypic data used in the analyses described herein.

Estimation of least square means and heritabilities: For the purpose of obtaining inbred line mean values adjusted for environmental effects, data were analyzed with a mixed model considering lines as fixed effects and environment, replication within environment, and line-by-environment interaction as random using Proc Mixed in SAS v9.2 (SAS Institute 2000-2004). Wald’s Z statistic was used to test the significance of each random factor in the model (Littell et al. 2006). Least squares means for lines were estimated from this mixed model and used as the input phenotype for association analysis. For the purpose of estimating heritability, a mixed model with all factors, including lines, as random effects was used.

Population Structure: Population structure can result in a systematic bias that produces false-positive associations if not accounted for in association analyses (Hirschhorn and Daly 2005). Population structure in this set of lines was previously analyzed using 89 SSR markers (Flint-Garcia et al. 2005). We re-analyzed the population structure using a subset of 5,000 SNP markers with no missing data and sampled from at least every 72 kbp interval of the maize physical map. STRUCTURE v2.3.3 software (Pritchard et al. 2000) was used to characterize the population structure of the maize association panel. The model implemented assumed that loci are independent within populations (Conrad et al. 2006; Falush et al. 2007), hence, the selection
of 5,000 markers used for the analysis was based on a relatively even distribution over the entire genome in which the smallest physical interval between any two markers used for the structure analysis was 72 kbp.

The method used to calculate population structure estimates the probability that a particular line belongs to a particular subpopulation \(Q_k\) given a fixed number of subpopulations \(k\) specified. Independent tests were conducted for \(k\) ranging from 1 to 12 using an admixture model, following a burn-in phase of \(1 \times 10^5\) and a sampling phase of \(5 \times 10^5\) replicates. Three runs were performed for each value of \(k\). By evaluating the change in model likelihood as \(k\) increased, we observed that initially, the likelihood increased monotonically as \(k\) increased, but after a point, the change in likelihood fluctuated slightly between increasing and decreasing values as \(k\) increased. We chose the optimal value of \(k\) as that value that produced the highest model likelihood before further increases in \(k\) resulted in a fluctuating response in likelihood to increasing \(k\) (Pritchard et al. 2000). Membership probabilities \(Q_k\) were used for assigning lines to subpopulations. Lines with highest membership probability \(Q_k\) less than 0.8 for all \(k\) were considered to result from admixture, hence, classified as “mixed”.

**Genotypic Correlation Analysis:** We estimated genotypic correlations among lesion mimic traits measured in this study and previously-derived quantitative resistance scores for three different diseases of maize measured on the same association panel but evaluated in different environment sets (Wisser et al. 2011): Southern leaf blight (SLB), Northern leaf blight (NLB), and Gray leaf spot (GLS). To reduce the impact of population structure on genotypic correlation estimates, we estimated correlations among inbred line residual values obtained after fitting population structure covariates \(\beta_k\) for each \(Q_k\) to least square means (for lesion mimic traits) or
best linear unbiased predictors (for disease scores) for each trait. We did not incorporate the realized genetic relationship matrix \((K)\) into the trait correlation estimation procedure because it is not appropriately scaled for variance-covariance component estimation (VanRaden, 2008; Zhang et al., 2009).

**Linkage disequilibrium analysis:** LD was quantified as \(r^2\) (Hill and Robertson 1968), and was estimated for all pairs of 47,445 SNPs using TASSEL v4.0 (Bradbury et al. 2007). We partitioned SNP pairs into those on the same chromosome (“linked” pairs) and those on different chromosomes (“unlinked” pairs). The 95th percentile \((Q_{95})\) of unlinked SNP LD \(r^2\) values was estimated from the distribution of values among all unlinked SNP pairs. We used this value as a threshold representing an upper bound of unlinked LD expected throughout the genome (Breseghello and Sorrels 2006). Within each chromosome, we classified SNP pairs according to physical distance into discrete distance ranges (e.g. 1 to 100 bp, 100 to 1000 bp, etc.), and estimated the distribution of linked LD \(r^2\) values for pairs within each distance class. All analyses besides generation of the \(r^2\) values were performed with R software (R Development Core Team 2005).

**Association analysis:** A matrix of genetic relationships between all pairs of lines \((K)\) was estimated using a subset of 4,000 SNPs. The markers used for the analysis were approximately uniformly distributed across the entire genome (the smallest physical interval between any two markers was 60 kbp) and had no missing data after excluding heterozygous SNP genotypes. The realized kinship coefficients were estimated in Tassel version 2.1 (Bradbury et al. 2007) using similarity based on marker identity by state (IBS). The similarity matrix was computed from the
distance matrix by subtracting all values from 2 then scaling so that the minimum value in the
matrix is 0 and the maximum value is 2. Tassel version 4.1.8 was used for the genome wide
association analysis based on a mixed linear model (MLM). The vector of phenotypes (y) was
modeled as:

\[ y = X\beta + Zu + e \]

where \( \beta \) represents a vector containing fixed effects, including the SNP marker being tested; \( u \)
represents a vector of random additive genetic effects associated with lines; \( e \) is a vector of
residual effects; while \( X \) and \( Z \) are incidence matrices relating \( y \) to \( \beta \) and \( u \), respectively. The
variances of the random effects are modeled as \( \text{Var}(u) = 2KV_g \), where \( K \) is \( n \times n \) matrix of
pairwise relative kinship coefficients defining degree of genetic covariance between lines and \( V_g \)
is the genetic variance (Yu et al. 2006).

The restricted maximum likelihood (REML) estimates of the variance components were
obtained using an efficient mixed-model association algorithm method (Kang et al. 2008)
(Zhang et al. 2010). The optimum compression mixed linear model and P3D options, which
increase statistical power and computational speed, were implemented by clustering individuals
into groups (Zhang et al. 2010). The p-values for each of the 47,445 tests of associations
between one SNP and one trait were used to estimate the positive false discovery rate (pFDR)
associated with each level of p value observed using the R package QVALUE version 1.0
(Storey and Tibshirani 2003).

Candidate gene selection
Genes located within or adjacent to associated SNPs were identified using the MaizeGDB genome browser (Andorf et al. 2010) or the maizeseq.org genome browser (Schnable et al. 2009). Annotations of the candidate genes were performed based on BLAST search of the amino acid sequence of the transcripts using the blastp (Altschul et al. 1997) and conserved domain search tools (Marchler-Bauer et al. 2005) on the NCBI website and the BLAST2GO software (Conesa et al. 2005).
Results

Heritability and analysis of variance

The Rp1-D21-H95 stock, which is heterozygous for the Rp1-D21 gene was crossed to a subset (231 lines) of the 302 line association panel and the resulting F₁ families were evaluated in replicated field trials over multiple environments for several traits associated with the severity of the auto-active HR phenotype conferred by the Rp1-D21 gene. The three field observation-derived traits (LES, HTR and SWR) all had high heritability, above 0.85 on a line-mean basis (Table S3). Of the image analysis-derived traits, only PCTLESAV and PCTLES4 had line mean heritability above 0.8. Line and line by environment interaction were significant contributors to variance for all traits (Table S3).

Correlation analysis

Genetic correlations were estimated between the field-derived HR-related traits and resistances to three different diseases of maize previously examined using the same association panel (Wisser et al. 2011): Southern leaf blight (SLB), Northern leaf blight (NLB), and Gray leaf spot (GLS). Correlation coefficients were estimated while taking population structure into account (Wisser et al. 2011). The traits measured on the Rp1-D21 population (LES, HTR and SWR) were highly genetically correlated with each other (|ρg| > 0.85, p < 0.001; Table 1). Correlations between these Rp1-D21-associated traits and the disease traits were moderately significant for only HTR and NLB (ρₕ = -0.11, p < 0.10) and LES and SLB (ρₕ = -0.12, p < 0.10).
Correlation coefficients estimated here among SLB, GLS and NLB resistance traits (0.52-0.59) were similar to those estimated from the previous study (0.55-0.67; Wisser et al. 2011) despite using a different marker data set for population structure estimation and a simplified approximate two-step estimation procedure in this study.

Assessment of population structure

Previous studies of similar samples of the same maize diversity panel employed 89 SSRs that detected 1694 alleles (Hamblin et al. 2007) and 94 SSRs that detected 2039 alleles (Liu et al. 2003) for estimating population structure. Population structure estimated here using 5,000 SNPs gave largely similar results to those reported previously (Figure S2, Table S1, S2). Compared to the previous analyses, some lines were reassigned from one of the three well-established maize germplasm groups (stiff-stalk (SS), non-stiff-stalk (NSS), or tropical-subtropical (TSS)) to the admixed group (containing lines with probability of membership in each of the three major germplasm groups < 0.8), but no lines were reassigned from one to another distinct population group. A large majority of the lines that were reassigned from one of the population groups to the mixed group in the current analysis had high probability of membership (p = 0.6-0.79) in their previously-assigned group (Table S1, S2); i.e., close to the arbitrary threshold used for group classification.

Population structure (Q) accounted for 16.5 % and 13.8 % of the variation in HTR and LES line means, respectively (Table S4). The realized kinship matrix captured most of the genotypic variance (77.1 % and 92.3 % for HTR and LES, respectively; Table S4).
Linkage disequilibrium in the diversity panel

We estimated the $r^2$ values of LD between each SNP and all other SNPs on different chromosomes (“unlinked SNP pairs”) to determine the empirical distribution of LD for unlinked SNPs. The 95th percentile ($Q_{95}$) of $r^2$ values for unlinked SNP pairs was estimated to be 0.04. We used this value as a threshold representing an upper bound of unlinked LD expected throughout the genome (Breseghello and Sorrells 2006). Genome-wide, considering SNPs on the same chromosome, mean LD $r^2$ dropped below 0.1 for SNP pairs separated by more than 10 kbp (Fig. 1). Mean LD $r^2$ for SNP pairs separated by more than 100 kbp was below the 0.04 threshold value defined for SNPs on different chromosomes.

Association mapping of loci modulating lesion mimic phenotype

Traits with a heritability above 0.8 on a line-mean basis were used for association analysis. The following traits met this criterion: lesion scores (LES), mutant to wild type height ratios (HTR), mutant to wild type stalk width ratio (SWR), percent necrotic lesions on the 3rd or 4th leaf (PCTLES4) and average of percent necrotic lesions (PCTLESAV). We performed association analysis using the least square mean values of inbred lines and a mixed linear model to adjust for background genetic relationships implemented in TASSEL version 4.1.8. Then we estimated the False Discovery Rate ($q$) for each SNP based on the empirical distribution of all SNP p-values for a given trait using the approach of Storey and Tibshirani (2003) (Figure, S3. S4). One SNP was associated with HTR at $q \leq 0.05$ and four additional SNPs associated with HTR at $q \leq 0.3$ (Fig. 2, Table 2). No SNPs had $q$-values below $q = 0.3$ for association tests with any of the other analysed traits. Among these traits however, analysis of LES yielded SNPs with
the lowest p-values. LES and HTR are highly correlated traits (Table 1) and all SNPs significantly associated with HTR were also found to be the most significant (lowest q-value) for associations with LES (Figure 2, Table 2).

To characterize local LD structure more accurately and in the genome regions surrounding the associations initially identified with the 50k Illumina Array, we re-scanned 2 Mbp windows surrounding each of these SNPs at higher marker density. This maize diversity panel was recently assayed for SNPs at more than 10-fold higher density using GBS. After re-scanning with the GBS data set, we detected a new strong SNP association at a locus on chromosome 10 (Table 2, Figure S5). The new SNP was 29,198 bp downstream of the initially identified SNP. These two SNPs are located within a block of relatively high LD (Figure S5), located between about 21,680,566 and 21,726,608 bp on chromosome 10. Furthermore, these two SNPs are in high LD with each other \( (\rho^2 = 0.36) \). We also detected one other SNP associated with HTR on chromosome 10. This third SNP is 100,526 and 129,724 bp from the other two significant SNPs, but despite its adjacent genomic location, is nearly in linkage equilibrium with them.

Candidate gene co-localized with associated SNPs

Using the filtered predicted gene set from the annotated maize genome based on maize inbred B73 (Schnable et al. 2009, http://www.maizesequence.org), we examined the genes that contained the SNPs that showed statistically significant associations with the traits. Several of these genes have predicted functions related to immune response pathways (Table 2), including a RING finger/U-Box domain-containing protein, a Nuclear encoded polymerase (NEP)
interacting-protein 2 (NIP2)/RING-H2 zinc finger domain containing protein, an Elongation
factor 1-alpha protein, a DNA polymerase alpha/epsilon subunit B protein, a heat shock 70 kDa
protein (HSP70) and a Ubiquitin E2 variant (UEV)/RING finger and WD (RWD) domain-
containing protein.

Allelic distribution at candidate genes

We estimated the frequency of alleles at the six SNPs significantly associated with HTR
in the three major maize germplasm groups (SS: stiff stalk; NSS: non stiff stalk and TS:
tropical/subtropical). Alleles enhancing the HR associated with Rp1-D21 are over-represented in
tropical/subtropical lines relative to other groups at all loci except the SNP on chromosome 9
(Table 3).
Discussion

In this study, we employed the MAGIC procedure (Johal et al. 2008) using F₁ families derived from crosses between a reference line with an allele conferring an auto-active HR phenotype, *Rp1-D21*, and a densely genotyped collection of 231 inbred lines to perform a GWAS. The goal of this strategy was to identify genomic variation that interacted epistatically with the *Rp1-D21* allele. These might include variation in defense response genes in pathways that are regulated by R-genes which are normally undetectable in a wild-type background. A shortcoming of the approach is that dominant alleles inherited from the reference line can mask functional variation harbored among the inbred lines. Although not implemented here, this can be addressed using different crossing schemes which allow for the detection of recessive alleles (Johal et al. 2008).

ANOVA and Heritability

The heritabilities of the field-observation derived traits, LES, HTR and SWR were all above 0.85 (Table S3). Of the image analysis derived traits, the PCTLES traits had heritabilities between 0.65 and 0.83 on a line mean basis but the heritabilities of other traits were much lower (Table S3). The main reasons for the lower heritabilities for the image analysis-derived traits likely included:

- A difference in the amount of data utilized. Image traits were calculated based on images from only two leaves per row. Field-observation scores were assessed on the entire row.

- The time period required to image the population. By necessity, images were captured over several days early in the season, a time of active growth when the plants were changing day to
day. LES was scored on a single day for the whole population at each time point. HTR and SWR were scored at the end of the season when the plants had stopped growing.

Correlation between disease resistance and defense response traits

The same maize association population had previously been assessed for resistance to three diseases SLB, GLS and NLB (Wisser et al. 2011). Strong correlations between resistances to these three diseases were identified, implying that the genetic mechanisms controlling these traits were partially shared. In order to determine whether some of the processes mediating the exaggerated defense response conferred by *Rp1-D21*, might also be involved in mediating disease resistance to SLB, GLS or NLB, we estimated the correlations between these traits measured in this population. Marginally significant correlations were observed between HTR and NLB (\(\rho_g = -0.11, \ p < 0.1\)) and between LES and SLB (\(\rho_g = -0.12, \ p < 0.1\)). While the HTR/NLB correlation was in the expected direction, (i.e a stronger *Rp1-D21* mediated defense response was associated with higher resistance), the LES/SLB correlation was not. Therefore it seems that variation affecting the strength of the maize HR was in large part unassociated with variation affecting resistance to SLB, NLB and GLS. Since HR is a mechanism associated predominantly with resistance to biotrophic pathogens and these three diseases are, to varying extents, necrotrophic (Jennings 1957), it could be argued that this result is not surprising.

Linkage disequilibrium in the maize association population

The selection of candidate genes using GWAS is based on the premise that a causative polymorphism will be in LD with markers in close proximity. The extent of LD determines
resolution i.e. the smaller the LD block, the better the resolution to detect causative SNPs/genes.

In this study we present a comprehensive genome-wide LD analysis of the maize genome. As found previously (Van Inghelandt et al. 2011; Yan et al. 2009), LD was somewhat variable across chromosomes and germplasm groups. On average, marker pairs separated by more than 10 kbp had an $r^2$ value less than 0.1 (Fig. 1). This level of LD is broadly in line with, though somewhat higher than previous estimates which were based on less extensive surveys of the genome. Remington et al. (2001) showed that LD around 6 genes in 102 inbred lines (a subset of the association population used here) generally declined rapidly, with $r^2$ values dropping below 0.1 within 1500 bp in most cases. A genome-wide LD scan of 327 loci in a population of 632 diverse inbred lines (which included the maize association population used here as well as other lines) showed that LD decay distances ranged between 1 and 10 kbp (Yan et al. 2009). Selection of candidate genes needs to be considered on a case by case basis since LD is highly variable across the genome (Figure 1, S5).

**False Discovery Rate estimation**

We used the approach of Storey and Tibshirani (2003) to estimate the false discovery rate (FDR q value) corresponding to each p-value obtained from GWAS. The relationship between FDR and p-values was estimated separately for each trait. This method attempts to estimate the proportion of true null hypotheses among all tests based on the observed distribution of p-values. If all null hypotheses (that the two alleles at each SNP have equal effects) were true, one expects an equal distribution of p-values across equally sized intervals from $p = 0$ to $p = 1$. If some proportion of null hypotheses are false, then one expects to observe a relatively constant proportion of tests with higher p-values (because these correspond to true null hypotheses) and
an inflated proportion of tests with p-values below some threshold, corresponding to a mixture of true null hypotheses and true false hypotheses. The method of Storey and Tibshirani (2003) estimates the proportion of truly null hypothesis based on the region of the p-value distribution that is approximately flat for the purpose of computing the expected false discovery rate corresponding to each p-value.

The two traits primarily studied here, HTR and LES, had high heritabilities, indicating strong genetic influence on the phenotypes, but the empirical distributions of GWAS p-values were skewed toward higher p-values for all traits (Figure S3, S4). Thus, we detected only a few significantly associated SNPs even at FDR of 0.30; the probability of false discoveries increased very rapidly to near one with only a small increase in p-values above the very lowest levels observed (Figure S3, S4). We expect that the remaining SNPs are truly null or have such small effects as to be undetectable with current sample sizes. These results suggest that many of the genes affecting these traits tend to have small effects, for which we have low power of detection due to a limited sample size and insufficient marker density for the low level of LD in the panel.

Influence of coancestry and population structure on statistical power of GWAS

We used the realized kinship matrix to minimize the chance of reporting false positive associations due to population structure or pedigree relationships among the lines of the diversity panel. The large amount of variation accounted for by the pairwise genetic relationships (Table S4) suggests that the inheritance of these traits is due primarily to additive polygenic effects. Power to detect individual SNP associations with the traits depends on the magnitude of their effects, their allele frequencies and their allelic distribution. In this case it is likely that several
SNPs that were associated with significant levels of variation were not detected since the effects of SNPs whose allelic distribution follows closely the background realized genetic relationships will contribute to the background additive genetic variance component modeled by the K matrix, and we will have low power to detect them in GWAS.

**Association analysis results**

Six SNPs were identified that were significantly associated with HTR (Table 2) three of these were located in a ∼130 kbp genomic region on chromosome 10 at 21,693,685 bp, 21,722,883 bp and 21,823,409 bp (which we will here call SNPs 1, 2 and 3 respectively). SNPs 1 and 2 are located in a region of high LD and are themselves in relatively high LD ($r^2 = 0.36$). Thus it is possible that SNPs 1 and 2 are associated with the same underlying causal variation. SNP3 however is in low LD with SNPs 1 and 2, suggesting that their associations reflect two distinct causal polymorphisms.

These chromosome 10 SNPs precisely co-localize with the *Hrml1* locus, a major QTL associated with variation in the same traits, which had been identified in an independent linkage analyses in several linkage mapping populations, most precisely in the advanced intercross line (AIL, *sensu* Darvasi and Soller 1995) IBM population which was derived from a cross between the inbreds B73 and Mo17 (Chaikam *et al.* 2011; Chintamanani *et al.* 2010). The present study provides a much higher resolution of the *Hrml1* region than before, narrowing the region of interest from ∼3 Mb potentially to single-gene resolution. The fact that we identified precisely the same QTL with several entirely independent data sets and two different analysis techniques validates this QTL and suggests that our data sets and analysis methods are robust and accurate.
The directions of the allelic effects were consistent between the IBM population QTL linkage analysis and this genome-wide association analysis (Table S5), as both the SNP and QTL allele that enhanced the $R_{p1}-D_{21}$ HR phenotype was carried by Mo17, and the SNP and QTL alleles that suppressed the HR phenotype were carried by B73. The large effect of the $Hr_{ml1}$ locus may therefore be explained in part because there appear to be two causal polymorphisms segregating together at this locus in the B73/Mo17 population. Similarly, the associated SNP on chromosome 9 is located close to a previously identified QTL interval in the IBM population (Chintamanani et al. 2010). This SNP is not polymorphic between B73 and Mo17 (Table S5).

Alleles enhancing the $R_{p1}-D_{21}$ HR phenotype, were over-represented in the tropical/subtropical germplasm group relative to the stiffstalk and non-stiffstalk groups (Table 3). An enhanced defense response would suggest higher disease resistance levels and agrees with observations that the tropical/subtropical germplasm group is in general more disease resistant than the other defined germplasms groups (Negeri et al. 2011; Wisser et al. 2011).

**Candidate genes**

We used the publically-available maize genome sequence to identify candidate genes encompassing or adjacent to these SNPs. Several of the candidate genes we identified play a role in the ubiquitination protein degradation pathway. In mammalian systems ubiquitin is critical for the regulation of several steps of the apoptosis pathway (Lee and Peter 2003). This was intriguing since both apoptosis and HR are forms of programmed cell death. Additionally, the plant ubiquitin pathway plays an important role in the plant defense response (Kadota et al. 2010; Peart et al. 2002). Ubiquitin ligation is a multi-step process that requires three classes of
enzymes (Ciechanover 1998), an E1 activating enzyme, a ubiquitin-conjugating enzyme E2 and finally an E3 ubiquitin-protein ligases.

Two of the identified candidate genes (on chromosomes 5 and 7, Table 2) contain RING-H2 finger domains, known to possess E3 ubiquitin-protein ligase activity and exhibit binding activity towards E2 ubiquitin-conjugating enzymes, mediating ubiquitination and degradation of the protein by the proteasome. The chromosome 5 associated SNP is within a gene that belongs to a class of E3 ligases defined by possession of a so-called U-box, a highly conserved ~70 amino acid modified RING-finger domain (Aravind and Koonin 2000; Koegl et al. 1999). Interestingly, U-box proteins appear to interact with molecular chaperones including HSP70 (Hatakeyama et al. 2004), another of our candidate genes (see below). The associated SNP on chromosome 5 creates a premature stop codon immediately downstream of the RING-finger domain. The chromosome 7 associated SNP is within a gene with strong homology to the nuclear-encoded plastid RNA polymerase (NEP) interacting-protein 2 (NIP2) which contains three transmembrane domains and one RING-H2 domain. The NIP2 gene has been implicated in pathogen defense response in *Nicotiana benthamiana* (Cheng et al. 2010).

The closest annotated gene to the associated SNP on chromosome 9 is predicted to be a eukaryotic elongation factor 1-alpha protein (EF1-α) gene, an evolutionarily conserved GTPase protein and part of the elongation factor-1 complex that catalyzes the enzymatic efficient delivery of charged tRNAs to the ribosome during protein elongation and has a critical role in translation fidelity and nuclear export of proteins (Uetsuki et al. 1989, Negrutskii and El'skaya 1998). A study by Talapatra et al. (2002) suggested that EF1-α expression conferred selective resistance to apoptosis induced by growth factor withdrawal and ER stress.
The other 3 associated SNPs were all located in the Hrml1 region on chromosome 10 as discussed above. SNP1 as SNP2 (as defined above) are in substantial LD with each other and define 2 candidate genes: SNP 1 is within a DNA polymerase alpha/epsilon subunit B gene and SNP2 is within an HSP70 gene. Although, due to LD it is difficult to tell precisely, which of these two genes is the more likely causative gene, based on functional annotation the HSP70 seems the better candidate. HSP70s are molecular chaperones, a component of the cell’s machinery involved in protein folding (Beere and Green 2001). The down-regulation of HSP70 has been shown to facilitate induction of apoptosis while its stress-induced up-regulation was shown to inhibit apoptosis in animal and plant cells (Cronjé et al. 2004; Parsell and Lindquist 1993). HSP70 was shown to be essential for HR associated with non-host resistance in tobacco (Kanzaki et al. 2003) and for basal resistance in Arabidopsis (Jelenska et al. 2010).

SNP3 on chromosome 10, (which is not in LD with other two chromosome 10 associated SNPs), is 2135 bp upstream of the start codon of a gene that has significant sequence similarity to an inactive form of E2 ubiquitin-conjugating enzyme predicted to be unable to catalyze ubiquitin transfer since it lacks the active cystine site. Nevertheless, the UEV domain has the ability to bind ubiquitin and may serve as a co-factor in ubiquitination reactions, as an ubiquitin sensor or to couple protein and ubiquitin binding functions to facilitate formation of multi-protein complexes (Pornillos et al. 2002; Teo et al. 2004). More recent studies (Spitzer et al. 2006) have annotated homologs of this maize gene in Arabidopsis as the ELC gene encoding Vps23p/TSG101 homolog, a key component of the ESCRT I-III machinery in yeast and animals that recognizes mono-ubiquitylated proteins and sorts them into the endosomal multivesicular body (MVB). The Arabidopsis ELC was shown to bind ubiquitin and localizes to endosomes, the
MVB, that contain numerous vesicles that are eventually fused with the vacuole/lysosome where proteins are degraded by luminal proteases (Odorizzi et al. 1998).

In conclusion, we have used the MAGIC approach combined with GWAS in the maize association panel as a powerful way to survey the maize allelic diversity to precisely map loci associated with natural variation in the HR defense response. In this way we identified six associated loci and a set of candidate genes that appear to be involved in connected functions controlling ubiquitination and programmed cell death. These novel findings would not have been possible using more conventional approaches such as mutational analyses or mapping of variation in the wild-type defense response.
Acknowledgements: This work was funded by USDA-ARS, Purdue University and an NSF Grant # 0822495. We thank Major Goodman and the Maize Genetics Cooperation Stock Center for donating seed and David Rhyne, Abbey Sutton, Joe Bundy, Ed Durren and Donna Stephens for help with fieldwork. We thank Ed Buckler, Jeff Glaubitz and the Maize Diversity project team for early access to the GBS genotyping dataset.
References


Liu, K., M. Goodman, S. Muse, J. S. Smith, E. Buckler et al., 2003 Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics **165**: 2117-2128.


Penning, B. W., G. S. Johal and M. M. McMullen, 2004 A major suppressor of cell death, slm1, modifies the expression of the maize (Zea mays L.) lesion mimic mutation les23. Genome 47: 961-969.


Dissection of maize kernel composition and starch production by candidate gene

Wisser, R. J., J. M. Kolkman, M. E. Patzoldt, J. B. Holland, J. Yu et al., 2011 Multivariate
analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a

Yan, J., T. Shah, M. L. Warburton, E. S. Buckler, M. D. McMullen et al., 2009 Genetic
characterization and linkage disequilibrium estimation of a global maize collection using

Yan, J., M. Warburton and J. Crouch, 2011 Association mapping for enhancing maize (Zea mays

Yu, J., and E. S. Buckler, 2006 Genetic association mapping and genome organization of maize.

method for association mapping that accounts for multiple levels of relatedness. Nature

Yu, L.-X., A. Lorenz, J. Rutkoski, R. Singh, S. Bhavani et al., 2011 Association mapping and
gene–gene interaction for stem rust resistance in CIMMYT spring wheat germplasm.

Zhang, Z., E. S. Buckler, T. M. Casstevens and P. J. Bradbury, 2009 Software engineering the
mixed model for genome-wide association studies on large samples. Briefings Bioinf. 10:
664-675.
Table Legends

Table 1: Genetic correlation coefficients between select traits and disease resistance score values obtained from a previous study (Wisser et al. 2011). Non-significant (ns, p > 0.1) correlation estimates not shown.

Table 2: Chromosomal locations, candidate genes and other parameters of the six SNPs identified as being significantly associated with height ratio (HTR) in this study.

Table 3: Allele frequencies of significantly associated SNPs in the maize germplasm groups (SS: stiff stalk; NSS: non stiff stalk and TS: tropical subtropical).

Figure Legends

Figure 1: Distribution of linkage disequilibrium measure ($r^2$) over various physical map distance classes between linked SNP marker pairs ($d$) over the entire maize genome. Horizontal dashed line indicates the $Q_{95}$ of the $r^2$ distribution between unlinked marker pairs (threshold value = 0.04) and an arbitrary fixed value of 0.1. The box-and-whiskers plot shows the smallest observation (lower whiskers), lower quartile (bottom part of box), median quartile (horizontal line in box), largest observation (sample maximum, upper whiskers) and the outliers (data points above upper whiskers). No. pairs represent the number of marker pairs in each distance class.

Figure 2: Results of GWAS showing the significant SNPs associations (arrows) with height ratio (HTR) (upper figure) and HR lesion severity (LES) (lower figure). The vertical axis indicates the $-\log^{10}$ of $p$-value scores and the horizontal axis indicates chromosomes and physical map positions of SNPs.
Figure 1: Distribution of linkage disequilibrium measure ($r^2$) over various physical map distance classes between linked SNP marker pairs ($d$) over the entire maize genome. Horizontal dashed line indicates the $Q_{95}$ of the $r^2$ distribution between unlinked marker pairs (threshold value = 0.04) and an arbitrary fixed value of 0.1. The box-and-whiskers plot shows the smallest observation (lower whiskers), lower quartile (bottom part of box), median quartile (horizontal line in box), largest observation (sample maximum, upper whiskers) and the outliers (data points above upper whiskers). No. pairs represent the number of marker pairs in each distance class.
**Figure 2**: Results of GWAS showing the significant SNPs associations (arrows) with height ratio (HTR) (upper figure) and HR lesion severity (LES) (lower figure). The vertical axis indicates the $-\log_{10}$ of \textit{p-value} scores and the horizontal axis indicates chromosomes and physical map positions.
Table 1. Genetic correlation coefficients between select traits and disease resistance score values obtained from a previous study (Wisser et al. 2011). Non-significant (ns, p > 0.1) correlation estimates not shown.

<table>
<thead>
<tr>
<th></th>
<th>HTR</th>
<th>SWR</th>
<th>SLB</th>
<th>GLS</th>
<th>NLB</th>
</tr>
</thead>
<tbody>
<tr>
<td>LES</td>
<td>-0.91**</td>
<td>-0.87**</td>
<td>-0.12*</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>HTR</td>
<td>0.85**</td>
<td>ns</td>
<td>ns</td>
<td>-0.11*</td>
<td></td>
</tr>
<tr>
<td>SWR</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLB</td>
<td></td>
<td>0.62**</td>
<td></td>
<td>0.67**</td>
<td></td>
</tr>
<tr>
<td>GLS</td>
<td></td>
<td></td>
<td></td>
<td>0.66**</td>
<td></td>
</tr>
</tbody>
</table>

Note: LES - Lesion score from field, HTR - Height ratio, SWR - Stalk width ratio, SLB- southern leaf blight resistance, GLS- grey leaf spot resistance, NLB- northern leaf blight resistance.

** p<0.01
* p<0.1
Table 2: Chromosomal locations, candidate genes and other parameters of the six SNPs identified as being significantly associated with height ratio (HTR) in this study.

<table>
<thead>
<tr>
<th>Chr</th>
<th>SNP physical position (bp)</th>
<th>HTR</th>
<th>LES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-value</td>
<td>Q-value</td>
<td>Allele</td>
</tr>
<tr>
<td>5</td>
<td>183,737,260</td>
<td>3.8 x 10⁻⁵</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>148,173,418</td>
<td>3.5 x 10⁻⁵</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>121,167,503</td>
<td>2.9 x 10⁻⁵</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
</tr>
<tr>
<td>10</td>
<td>21,693,685</td>
<td>3.3 x 10⁻⁷</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
</tr>
<tr>
<td>10</td>
<td>21,722,883</td>
<td>4.1 x 10⁻⁶</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>10</td>
<td>21,823,409</td>
<td>8.7 x 10⁻⁵</td>
<td>0.182</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
</tr>
</tbody>
</table>

¹Chr: chromosome; ²Alleles are from homozygote genotypes; ³N: total number of lines with the specific SNP genotype; ⁴positive allelic effects for HTR and LES imply a suppressive and enhancing effect on the HR phenotype respectively; ⁵R²: proportion of phenotypic variance explained by SNP; ⁶based on SNPs from illumina chip; ⁷based on SNPs obtained by genotyping-by-sequencing (GBS).
Table 3: Allele frequencies of significantly associated SNPs in the maize germplasm groups (SS: stiff stalk; NSS: non stiff stalk and TS: tropical subtropical).

<table>
<thead>
<tr>
<th>Chr</th>
<th>SNP physical position (bp)</th>
<th>Allele increasing HR 2</th>
<th>Allele frequency(%) 3</th>
<th>p-value 4</th>
<th>N 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>183,737,260</td>
<td>G</td>
<td>83.3</td>
<td>41.5</td>
<td>93.8</td>
</tr>
<tr>
<td>7</td>
<td>148173418</td>
<td>A</td>
<td>0.0</td>
<td>14.6</td>
<td>29.0</td>
</tr>
<tr>
<td>9</td>
<td>121167503</td>
<td>G</td>
<td>64.7</td>
<td>5.1</td>
<td>25.9</td>
</tr>
<tr>
<td>10</td>
<td>21,693,685</td>
<td>G</td>
<td>38.9</td>
<td>63.4</td>
<td>81.3</td>
</tr>
<tr>
<td>10</td>
<td>21,722,883</td>
<td>T</td>
<td>33.3</td>
<td>71.8</td>
<td>86.7</td>
</tr>
<tr>
<td>10</td>
<td>21,823,409</td>
<td>C</td>
<td>33.3</td>
<td>53.8</td>
<td>76.7</td>
</tr>
</tbody>
</table>

1 Chr: chromosome; 2 Alleles that increase hypersensitive response in the LES (visual lesion score) and HTR (mutant: wild type ratio) traits; 3 Alleles are from homozygote genotypes; 4 p-values after testing the null that the proportions (probabilities of success) in sub-populations are the same (prop.test in R software); 5 N: total number of lines included in analysis.