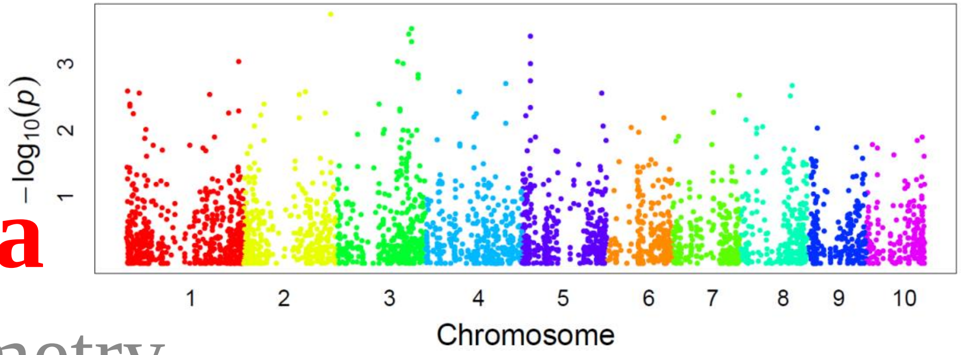
Implementing genomic selection and comparing it to marker-assisted selection



Alexander E. Lipka

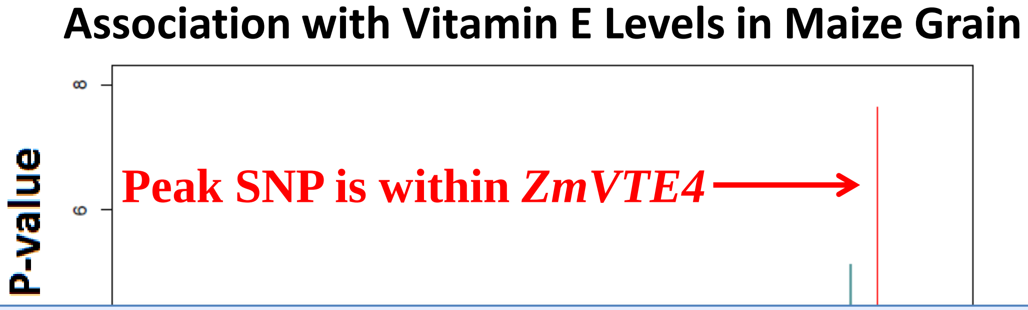
Assistant Professor of Biometry

Department of Crop Sciences

University of Illinois

USA

Genome-wide association study (GWAS)


Association with Vitamin E Levels in Maize Grain

Peak SNP is within *ZmVTE4*

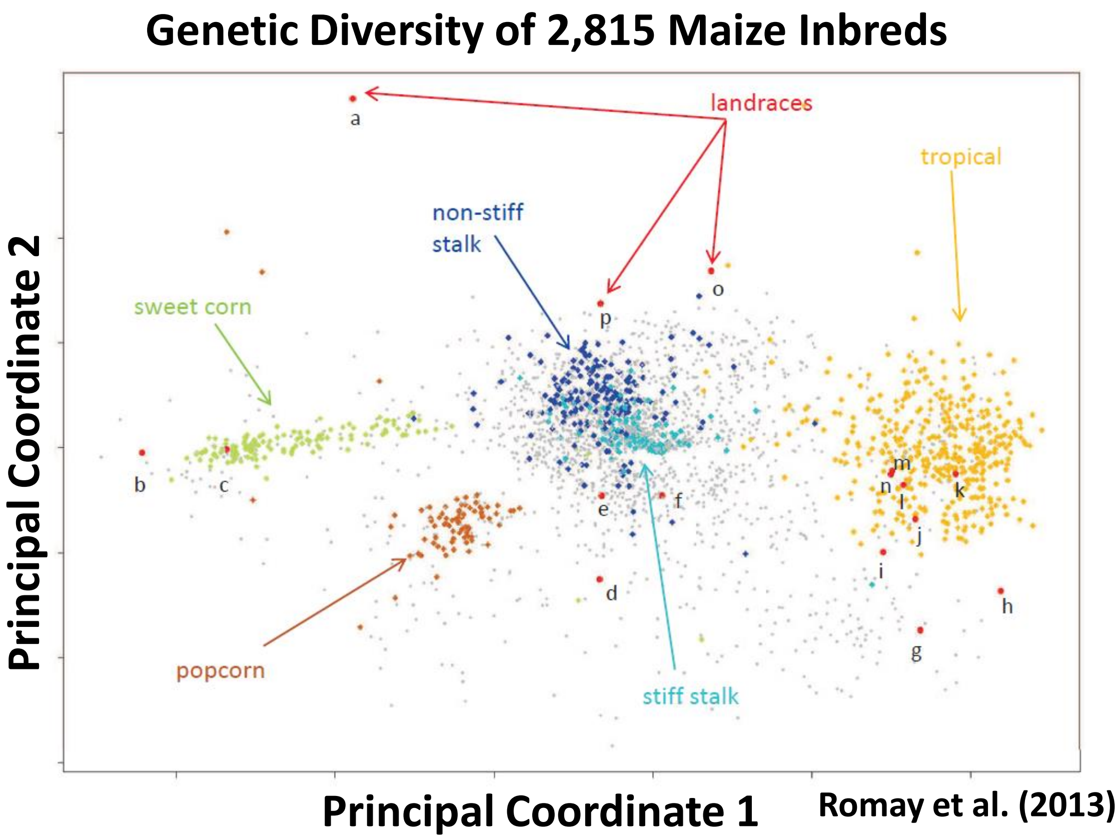
Markers exhibiting peak associations with traits are potential targets for marker-assisted selection (MAS)

- Identify genomic regions associated with a phenotype
- Fit a statistical model at each SNP in genome
- Use fitted models to test H_0: No association with SNP and phenotype

Examples of GWAS identifying potential targets for MAS breeding efforts

- Rincker et al. (2016): Targets for brown stem rot resistance in soybean
- Lipka et al. (2013): Targets for boosting vitamin E and antioxidant levels in maize grain
- Owens/Lipka et al. (2014): Targets for boosting provitamin A and other carotenoid levels in maize grain

Genetic diversity can lead to false positives in a GWAS



Romay et al. (2013)

- Two sources for false positives:
  - – Population Structure
  - – Familial Relatedness

Mixed models reduce false positives in GWAS

Random effects: account for familial relatedness

Grand Mean

Marker effect

$$Y_i = \mu + \sum_{j=1}^{3} \beta_j PC_{ji} + \alpha x_i + Line_i + \varepsilon_i$$

Phenotype of i^{th} individual

Fixed effects: account for population structure

Observed SNP alleles of i^{th} individual

Random error term

- $(Line_1, \ldots, Line_n) \sim \text{MVN}(\mathbf{0},\ 2K\sigma_G^2)$
- K = kinship matrix
- $\varepsilon_i \sim$ i.i.d. N(0, σ_E^2)

Measures relatedness between individuals

Yu et al. (2006)

Computational approaches for reducing computational burden

- The unified mixed linear model is a common

GAPIT R package (Lipka et al. 2012):

- **Employs computationally-efficient approaches for GWAS**
- **Makes it possible to perform mixed-model GWAS on an ordinary computer**

  – Newly-developed model fitting approaches need to be used to address this challenge

Unified mixed linear model (MLM)

Random effects: account for familial relatedness

Grand Mean

Marker effect

$$Y = \mu + \sum^{3} \beta PC + \alpha x + Line + \varepsilon$$

- Variance component estimation is computationally intensive
- GAPIT employs two approaches to reduce this computational burden

- $(Line_1, \ldots, Line_n) \sim MVN(0, 2K\sigma^2_G)$
- K = kinship matrix
- $\varepsilon_i \sim$ i.i.d. $N(0, \sigma^2_E)$

Measures relatedness between individuals

Yu et al. (2006)

Approach 1: Compressed mixed linear model

$$Y_i = \mu + \sum_{j=1}^{3} \beta_j PC_{ji} + \alpha x_i + Group_i + \varepsilon_i$$

- **Reduces computational time because it works with a smaller kinship matrix**

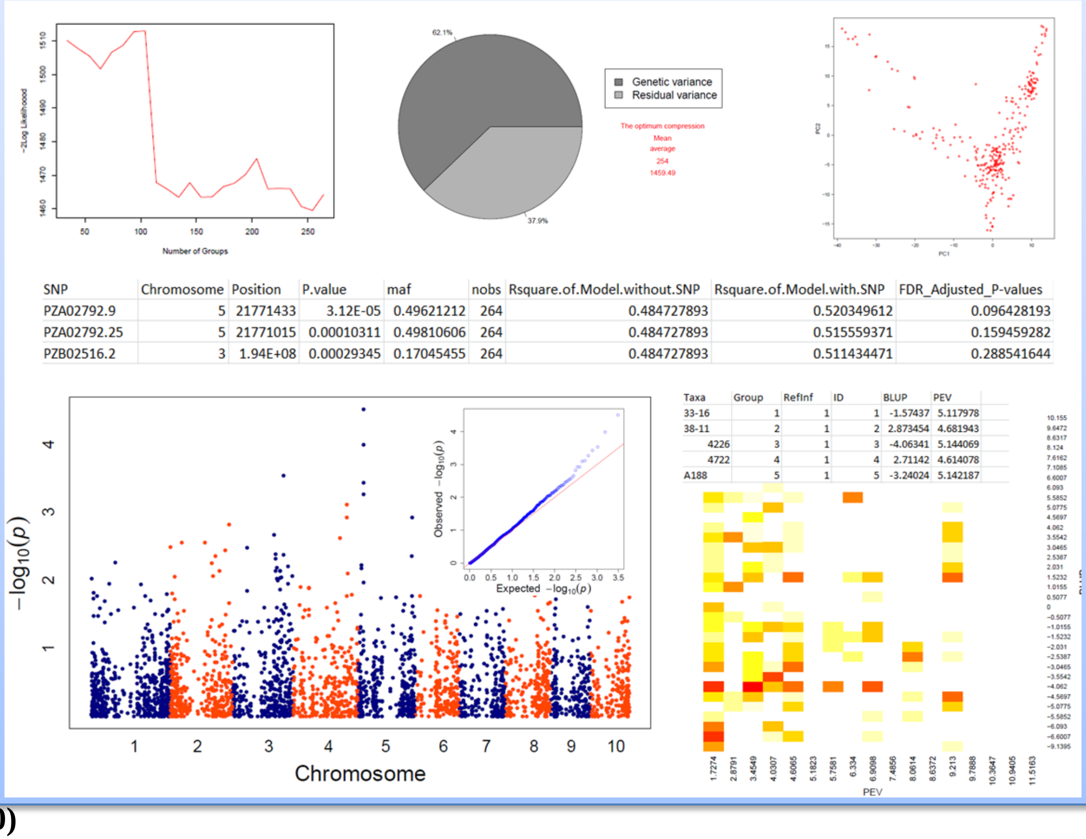
using kinship matrix

- $(Line_1, \ldots, Line_n) \sim$ MVN(0, 2K
- $(Group_1, \ldots, Group_k) \sim$ MVN(0, $2K_C\sigma_G^2$)
- K = kinship matrix
- K_C = group (“compressed”) kinship matrix
- $\varepsilon_i \sim$ i.i.d. N(0, σ_E^2)

Zhang et al. (2010)

Approach 2: Population parameters previously determined (P3D)

Output Summary



SNP	Chromosome	Position	P.value	maf	nobs	Rsquare.of.Model.without.SNP	Rsquare.of.Model.with.SNP	FDR_Adjusted_P-values
PZA02792.9	5	21771433	3.12E-05	0.49621212	264	0.484727893	0.520349612	0.096428193
PZA02792.25	5	21771015	0.00010311	0.49810606	264	0.484727893	0.515559371	0.159459282
PZB02516.2	3	1.94E+08	0.00029345	0.17045455	264	0.484727893	0.511434471	0.288541644

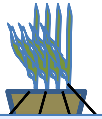
Taxa	Group	RefInf	ID	BLUP	PEV
33-16	1	1	1	-1.57437	5.117978
38-11	2	1	2	2.873454	4.681943
4226	3	1	3	-4.06341	5.144069
4722	4	1	4	2.71142	4.614078
A188	5	1	5	-3.24024	5.142187

Lipka et al. (2012)

Zhang et al. (2010)

Accounting for multiple reps and locations

Seeds obtained from a germplasm bank



- Fit a mixed model accounting for genetic, environmental, and genetic x environmental (GxE) sources of trait variation
- Output from this model:

- BLUPs/BLUEs trait values for each taxa
- Estimates of trait variation attributable to each source

multiple reps and locations?

Statistical model used to obtain best linear unbiased predictions (BLUPs)

Grand Mean

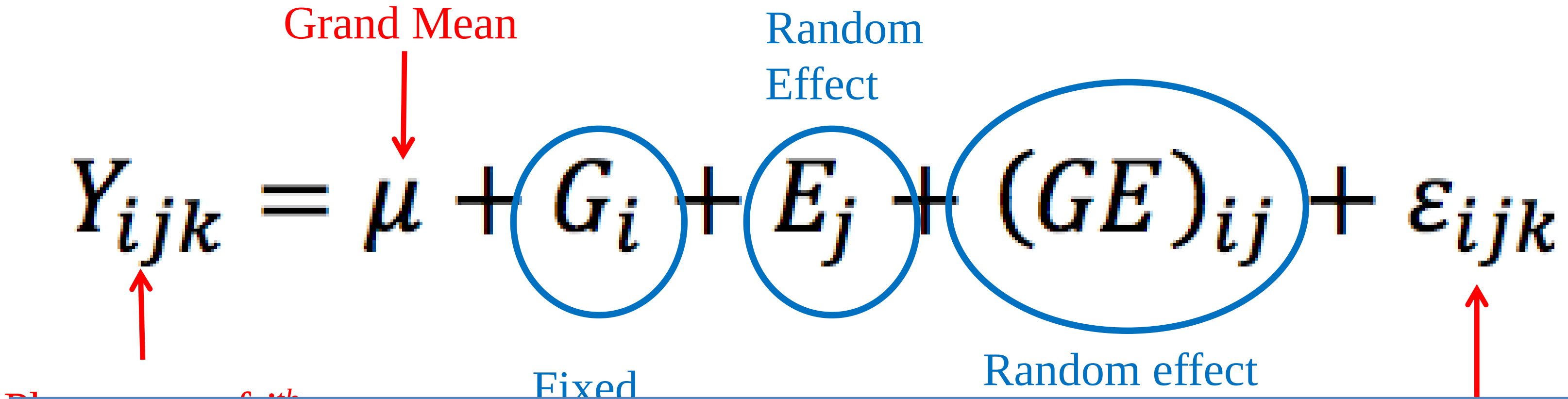
Random Effect

$$Y_{ijk} = \mu + G_i + E_j + (GE)_{ij} + \varepsilon_{ijk}$$

- **Output 1: BLUPs of the genotype effect**
- **Output 2: Variance component estimates for calculating heritabilities**

- G_i = Random Genotype Effect
- E_i = Random Environment Effect
- $(GE)_{ij}$ = Random Genotype x Environment Effect

Statistical model used to obtain best linear unbiased estimators (BLUEs)



Grand Mean

Random Effect

$$Y_{ijk} = \mu + G_i + E_j + (GE)_{ij} + \varepsilon_{ijk}$$

Fixed

Random effect

- **Output: BLUEs of the genotype effect**

term

- G_i = Fixed Genotype Effect
- E_i = Random Environment Effect
- $(GE)_{ij}$ = Random Genotype x Environment Effect

BLUPs vs BLUEs

- BLUPs:
  - Advantage: Makes more sense from a biological perspective
  - Disadvantage 1: BLUPs “shrink” values towards the mean
  - Disadvantage 2: Fitting random effects is more computationally intensive than fitting fixed effects
- BLUEs:
  - Advantage 1: BLUEs do not shrink values towards the mean
  - Advantage 2: Less computationally intensive
  - Disadvantage: Makes less sense from a biological perspective

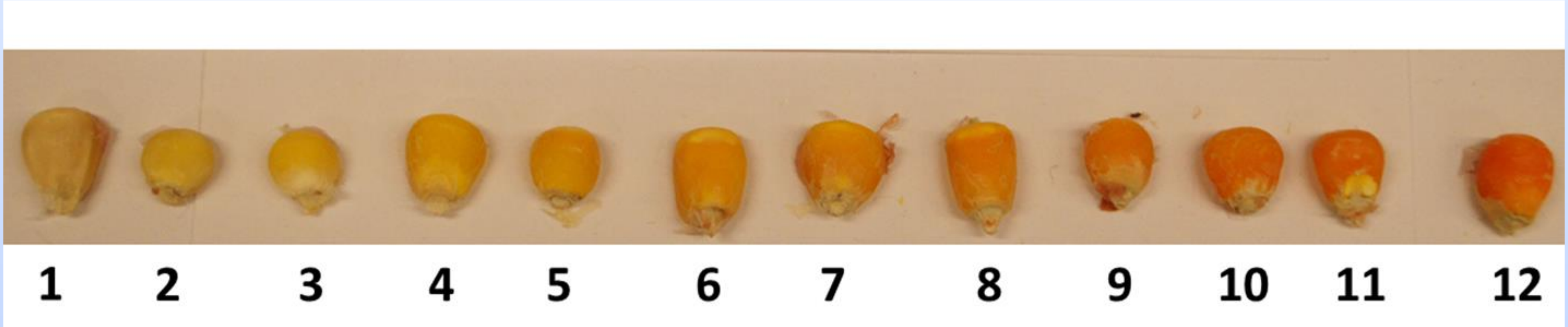
BLUPs and BLUEs: Some Technical Notes

- In plant breeding, estimate of grand mean is added to BLUPs and BLUEs
  - Rationale: BLUEs/BLUPs will be in the same units of measurement as raw trait data
  - After adding grand mean estimate, they are still called BLUPs/BLUEs
- Consider transforming your phenotypic data before fitting statistical models:
  - Rationale: This would help with deviations from normality and constant variance assumptions

Software I used to obtain BLUPs and BLUEs

- SAS:
  - Advantage: (Relatively) simple to use
  - Disadvantage 1: Annual license fee
  - Disadvantage 2: Takes a long time to compute
- ASReml:
  - Advantage: Can fit very complicated models quickly
  - Disadvantage 1: Not simple to use
  - Disadvantage 2: Expensive annual license fee
- R:
  - Advantage: Free
  - Disadvantage: Potentially not as extensively tested as SAS and ASREML

Phenotype: kernel color visually assessed using standardized color scale

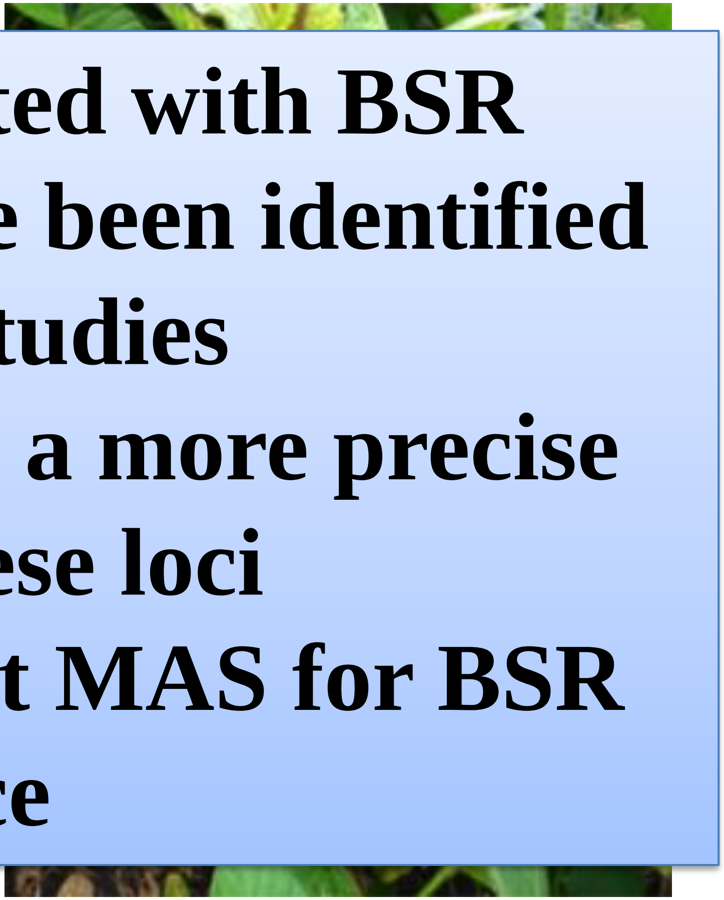


- Also included AR1xAR1 correlation structure to account for spatial variation
- Backwards elimination conducted to remove non-significant effects
- Analysis conducted in ASREML

Chandler/Lipka et al. (2013)

Example: Rincker et al. (2016)

- Brown stem rot (BSR) and



- **Three genes associated with BSR resistance, *Rbs1-3*, have been identified in previous studies**
- **Critical need to obtain a more precise location of these loci**
- **Result in more efficient MAS for BSR resistance**

Source: cornandsoybeandigest.com/

Separate GWAS performed on four association panels

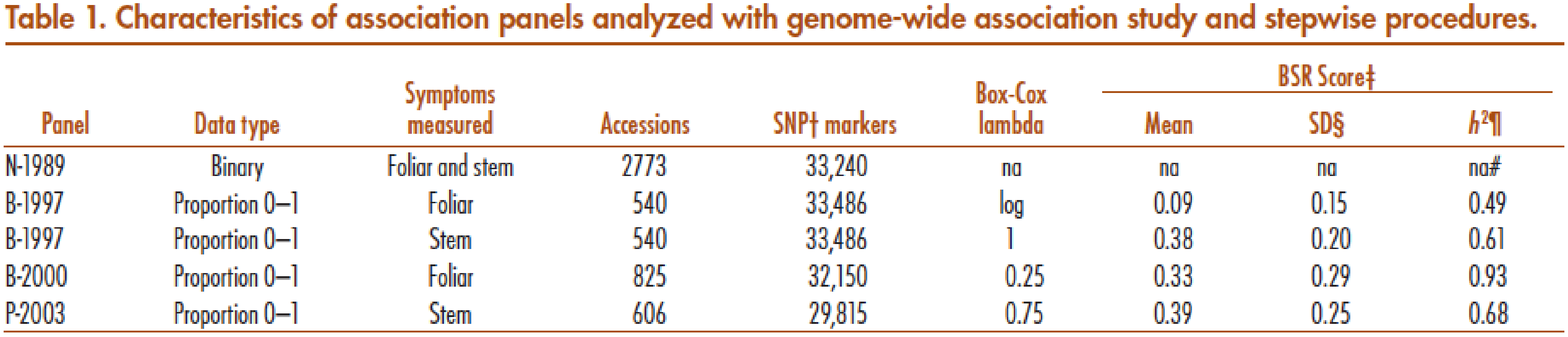


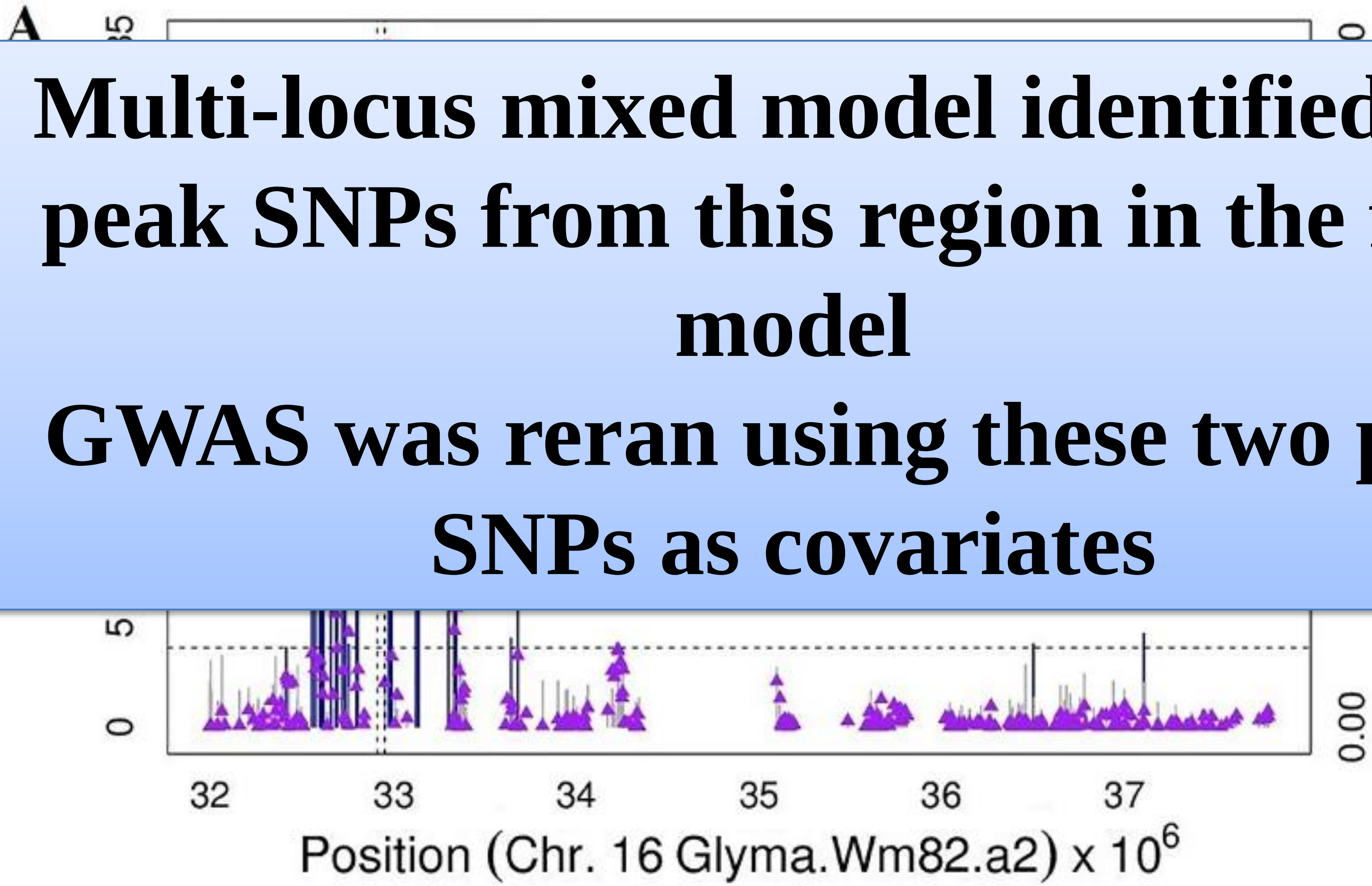
Table 1. Characteristics of association panels analyzed with genome-wide association study and stepwise procedures.

Panel	Data type	Symptoms measured	Accessions	SNP† markers	Box-Cox lambda	BSR Score‡		
						Mean	SD§	h^2¶
N-1989	Binary	Foliar and stem	2773	33,240	na	na	na	na#
B-1997	Proportion 0–1	Foliar	540	33,486	log	0.09	0.15	0.49
B-1997	Proportion 0–1	Stem	540	33,486	1	0.38	0.20	0.61
B-2000	Proportion 0–1	Foliar	825	32,150	0.25	0.33	0.29	0.93
P-2003	Proportion 0–1	Stem	606	29,815	0.75	0.39	0.25	0.68

- N-1989 panel:
  - – Binary phenotype: logistic regression + stepwise model selection
- Other panels:
  - – Quantitative phenotype: Unified MLM + multi-locus mixed model

Rincker et al. (2016)

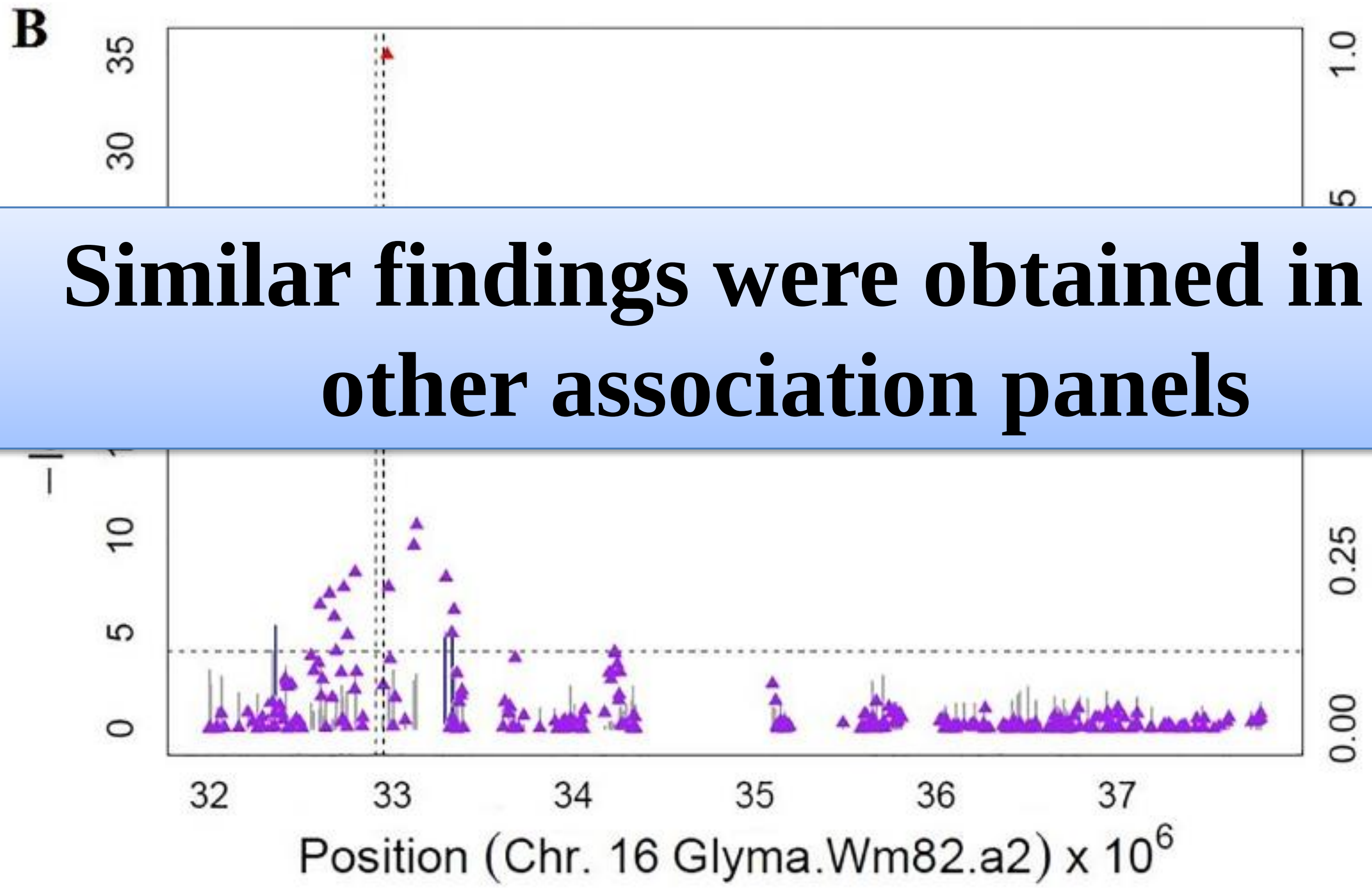
Unified MLM GWAS identifies signals near *Rbs1-Rbs3*



- **Multi-locus mixed model identified two peak SNPs from this region in the final model**
- **GWAS was reran using these two peak SNPs as covariates**

Rincker et al. (2016)

Peak SNPs from MLMM reduces explains most of *Rbs1-Rbs3* signal



- **Similar findings were obtained in the other association panels**

Rincker et al. (2016)

Breeding Ramifications

Source: blogs.ext.vt.edu

- Previous *Rbs1-Rbs3* signals been refined to a 0.3 Mb region on Chromosome 16
- Should facilitate both MAS-based approaches and gene cloning efforts
- Demonstrates the utility of GWAS in soybean

Rincker et al. (2016)

Biofortification

- Identify target genes associated with nutrients in crops

Source: www. aboutharvest.com

- Increase nutritional value of local crop varieties by selecting on these target genes

- Results in increased availability of essential nutrients

Compounds analyzed in Lipka et al. (2013)

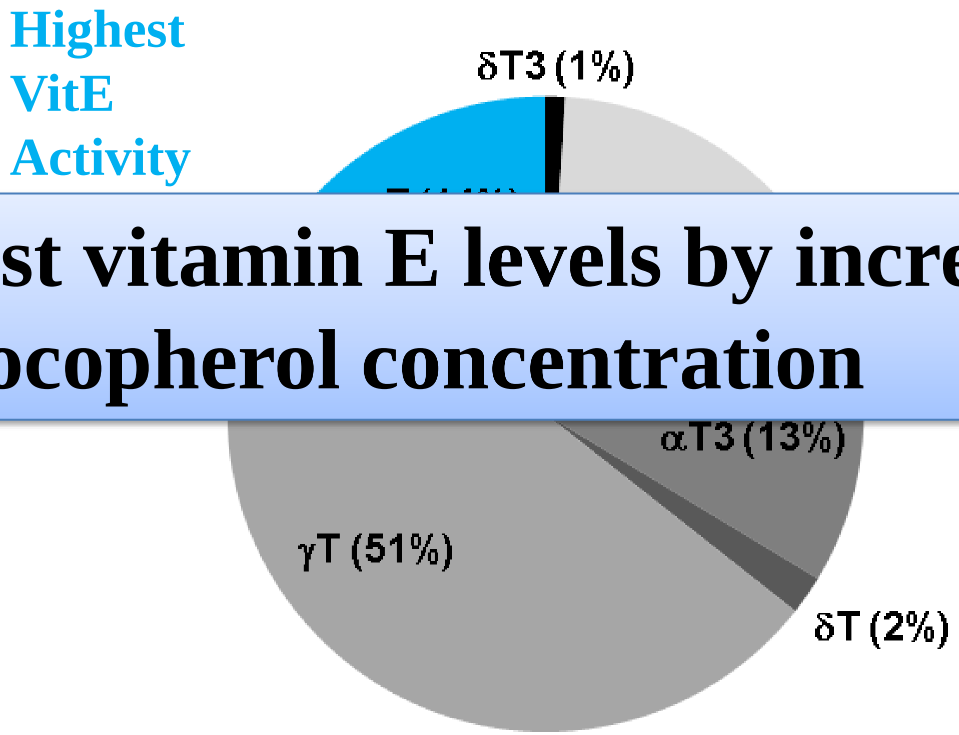
Source: wartremovalexperts.com

- **Tocochromanols**
  - Lipid-soluble antioxidants
  - Consist of **tocopherols (T)** and **tocotrienols (T)**
  - **α-tocopherol (αT)** has greatest vitamin E activity

- **Vitamin E**
  - Essential nutrient
  - Suboptimal dietary intake exists in specific population segments
  - Deficiency associated with cardiovascular disease and decreased immune function

Grain tocochromanol compositions across a maize diversity panel

Distribution of
Tocochromanol Compounds



- Boost vitamin E levels by increasing α-tocopherol concentration

Data analyzed in Lipka et al. (2013)

Source: Brenda Owens

- 281-member Goodman diversity panel
- Grown at Purdue University in 2009 and 2010 field seasons
- Compound levels quantified in grain:
  - Tocochromanols for 252 lines

Phenotypic data used for analysis

- 20 tocochromanol compounds, sums, ratios, and proportions were analyzed in GAPIT
- GWAS was conducted using 294,092 SNPs with minor allele frequency $\geq$ 0.05

Mixed model accounting for field season effects fitted to each phenotype

- Best linear unbiased predictors (BLUPs) of lines from each model used as phenotypes for our GWAS

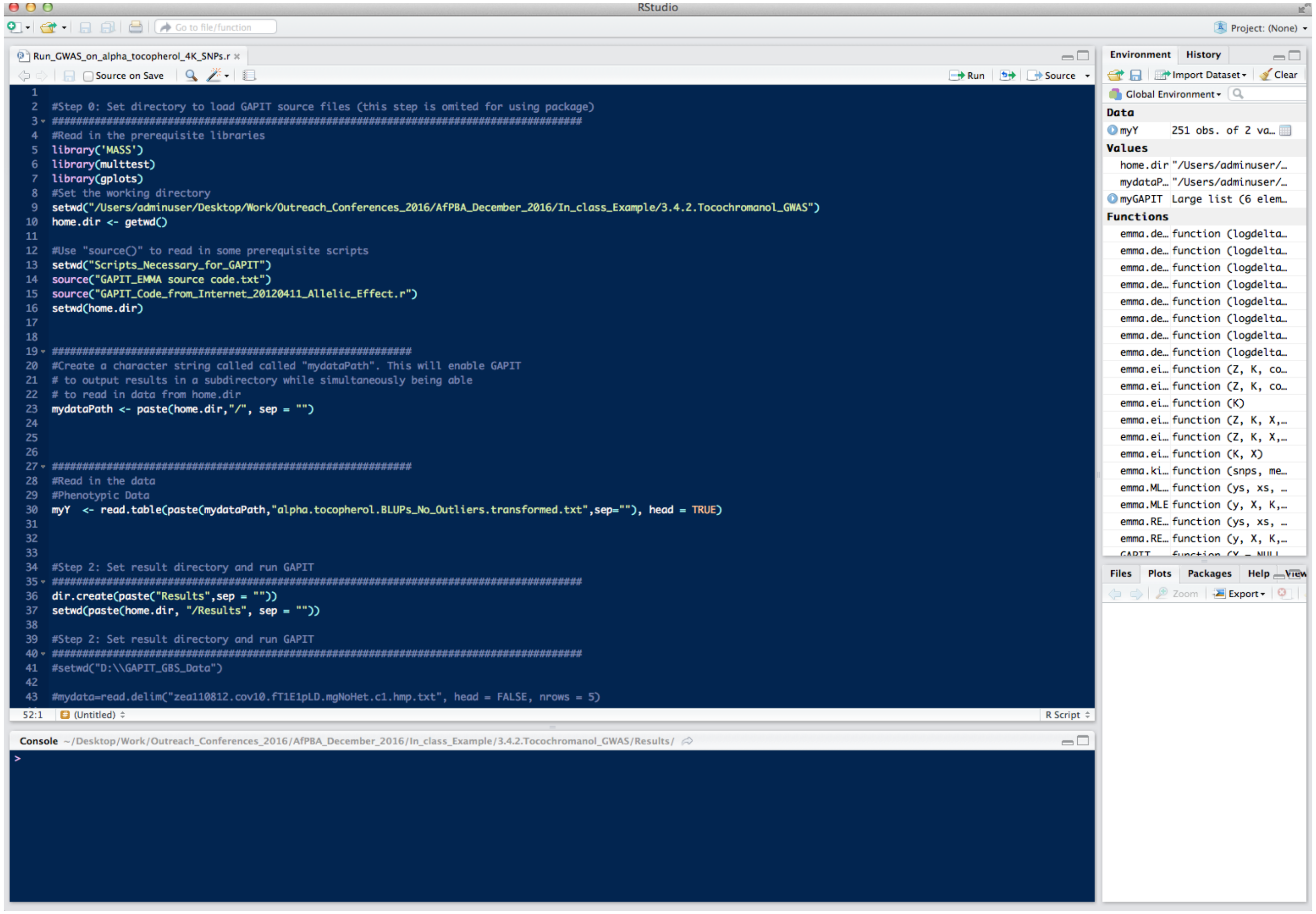
In-class example: GWAS scan of Lipka et al. (2013) data subset

- Trait: α-tocopherol
  - Has the greatest Vitamin E activity
- Marker subset:
  - 3,093 marker set obtained from various marker technologies (i.e., the 4k marker set)
- GWAS software used: Genome association and prediction integrated tool (GAPIT)
  - Unified mixed linear model is fitted at each SNP
  - Population parameters previously determined (P3D) used to save computational time

In-class example: GWAS scan of Lipka et al. (2013) data subset

- 4K_SNPsmdp_genotype_test1_GBS_Names1.hmp.txt
  - Genotypic data: 3,093 SNPs
- alpha.tocopherol.BLUPs_No_Outliers.transformed.txt
  - Phenotypic data: α-tocochperhol levels
- Scripts_Necessary_for_GAPIT
  - Folder containing scripts to be read into R
- Run_GWAS_on_alpha_tocopherol_4K_SNPs.r
  - R script for conducting the GWAS

In-class example: GWAS scan of Lipka et al. (2013) data subset

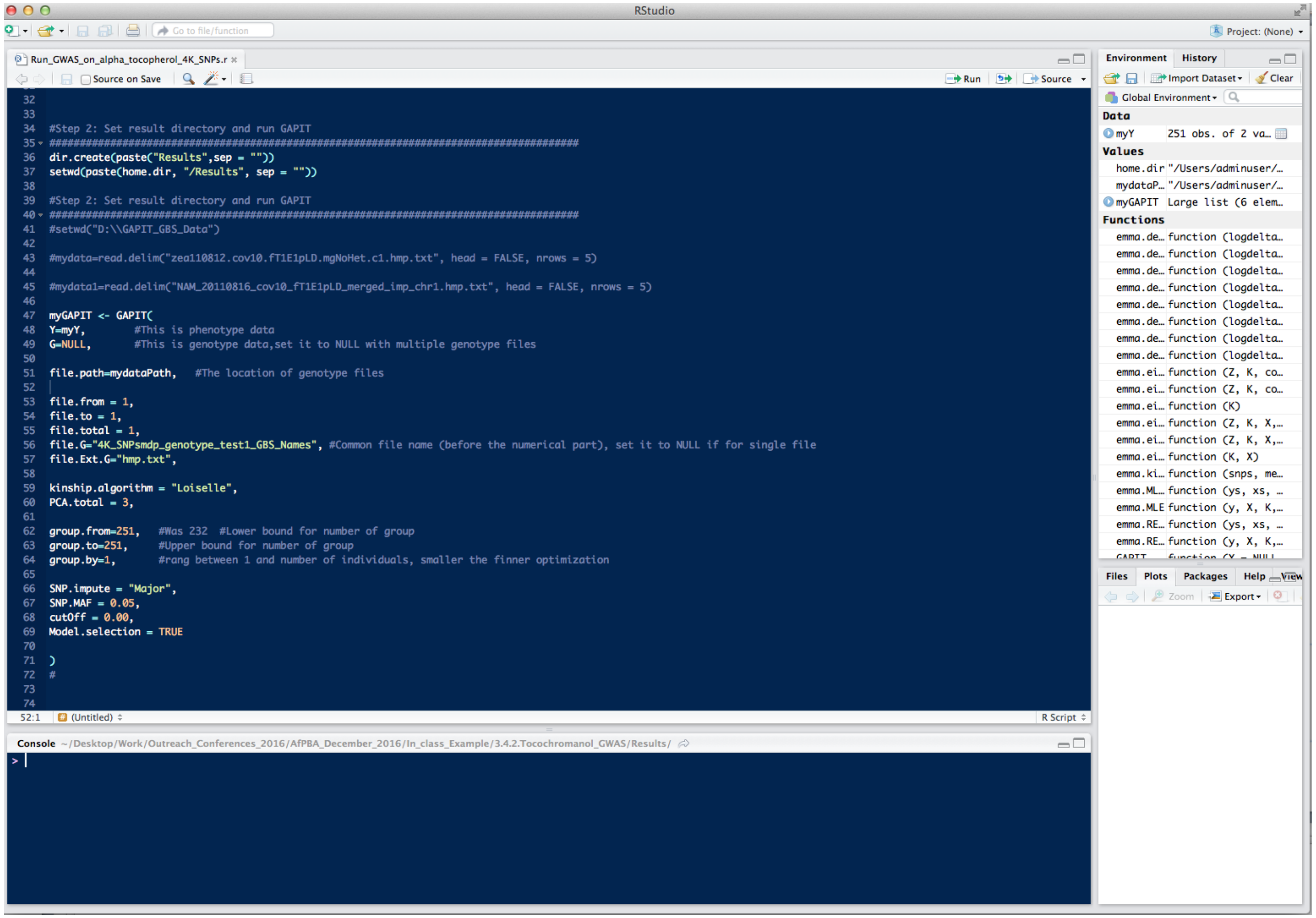


```r

#Step 0: Set directory to load GAPIT source files (this step is omited for using package)
#######################################################################################
#Read in the prerequisite libraries
library('MASS')
library(multtest)
library(gplots)
#Set the working directory
setwd("/Users/adminuser/Desktop/Work/Outreach_Conferences_2016/AfPBA_December_2016/In_class_Example/3.4.2.Tocochromanol_GWAS")
home.dir <- getwd()

#Use "source()" to read in some prerequisite scripts
setwd("Scripts_Necessary_for_GAPIT")
source("GAPIT_EMMA source code.txt")
source("GAPIT_Code_from_Internet_20120411_Allelic_Effect.r")
setwd(home.dir)


###########################################################
#Create a character string called called "mydataPath". This will enable GAPIT
# to output results in a subdirectory while simultaneously being able
# to read in data from home.dir
mydataPath <- paste(home.dir,"/", sep = "")



###########################################################
#Read in the data
#Phenotypic Data
myY  <- read.table(paste(mydataPath,"alpha.tocopherol.BLUPs_No_Outliers.transformed.txt",sep=""), head = TRUE)



#Step 2: Set result directory and run GAPIT
#######################################################################################
dir.create(paste("Results",sep = ""))
setwd(paste(home.dir, "/Results", sep = ""))

#Step 2: Set result directory and run GAPIT
#######################################################################################
#setwd("D:\\GAPIT_GBS_Data")

#mydata=read.delim("zea110812.cov10.fT1E1pLD.mgNoHet.c1.hmp.txt", head = FALSE, nrows = 5)
```

```
Console ~/Desktop/Work/Outreach_Conferences_2016/AfPBA_December_2016/In_class_Example/3.4.2.Tocochromanol_GWAS/Results/
>
```

```
Data
myY       251 obs. of 2 va…
Values
home.dir  "/Users/adminuser/…
mydataP…  "/Users/adminuser/…
myGAPIT   Large list (6 elem…
Functions
emma.de…  function (logdelta…
emma.de…  function (logdelta…
emma.de…  function (logdelta…
emma.de…  function (logdelta…
emma.de…  function (logdelta…
emma.de…  function (logdelta…
emma.de…  function (logdelta…
emma.de…  function (logdelta…
emma.ei…  function (Z, K, co…
emma.ei…  function (Z, K, co…
emma.ei…  function (K)
emma.ei…  function (Z, K, X,…
emma.ei…  function (Z, K, X,…
emma.ei…  function (K, X)
emma.ki…  function (snps, me…
emma.ML…  function (ys, xs, …
emma.MLE  function (y, X, K,…
emma.RE…  function (ys, xs, …
emma.RE…  function (y, X, K,…
GAPIT     function (Y = NULL
```

In-class example: GWAS scan of Lipka et al. (2013) data subset



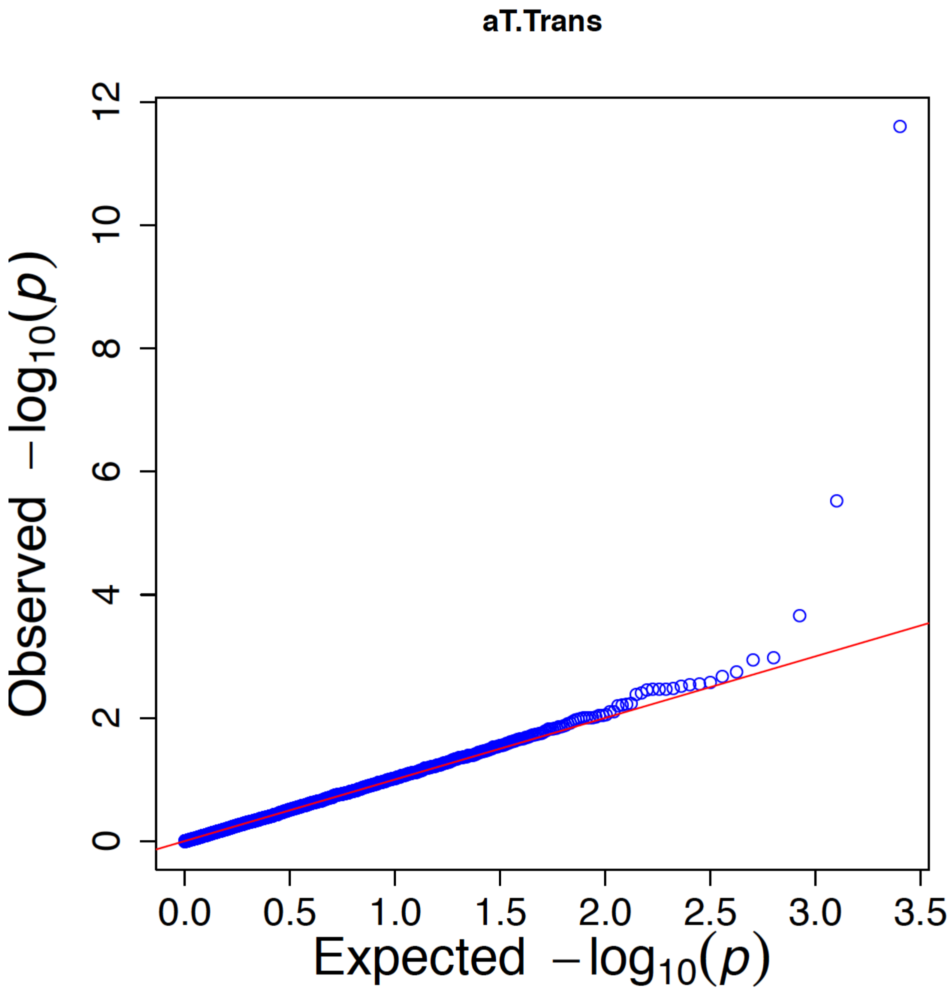
```r


#Step 2: Set result directory and run GAPIT
#########################################################################################
dir.create(paste("Results",sep = ""))
setwd(paste(home.dir, "/Results", sep = ""))

#Step 2: Set result directory and run GAPIT
#########################################################################################
#setwd("D:\\GAPIT_GBS_Data")

#mydata=read.delim("zea110812.cov10.fT1E1pLD.mgNoHet.c1.hmp.txt", head = FALSE, nrows = 5)

#mydata1=read.delim("NAM_20110816_cov10_fT1E1pLD_merged_imp_chr1.hmp.txt", head = FALSE, nrows = 5)

myGAPIT <- GAPIT(
Y=myY,          #This is phenotype data
G=NULL,         #This is genotype data,set it to NULL with multiple genotype files

file.path=mydataPath,   #The location of genotype files

file.from = 1,
file.to = 1,
file.total = 1,
file.G="4K_SNPsmdp_genotype_test1_GBS_Names", #Common file name (before the numerical part), set it to NULL if for single file
file.Ext.G="hmp.txt",

kinship.algorithm = "Loiselle",
PCA.total = 3,

group.from=251,   #Was 232  #Lower bound for number of group
group.to=251,     #Upper bound for number of group
group.by=1,       #rang between 1 and number of individuals, smaller the finner optimization

SNP.impute = "Major",
SNP.MAF = 0.05,
cutOff = 0.00,
Model.selection = TRUE

)
#


```

```
Data
myY        251 obs. of 2 va…
Values
home.dir   "/Users/adminuser/…
mydataP…   "/Users/adminuser/…
myGAPIT    Large list (6 elem…
Functions
emma.de…   function (logdelta…
emma.de…   function (logdelta…
emma.de…   function (logdelta…
emma.de…   function (logdelta…
emma.de…   function (logdelta…
emma.de…   function (logdelta…
emma.de…   function (logdelta…
emma.de…   function (logdelta…
emma.ei…   function (Z, K, co…
emma.ei…   function (Z, K, co…
emma.ei…   function (K)
emma.ei…   function (Z, K, X,…
emma.ei…   function (Z, K, X,…
emma.ei…   function (K, X)
emma.ki…   function (snps, me…
emma.ML…   function (ys, xs, …
emma.MLE   function (y, X, K,…
emma.RE…   function (ys, xs, …
emma.RE…   function (y, X, K,…
GAPIT      function (Y = NULL
```

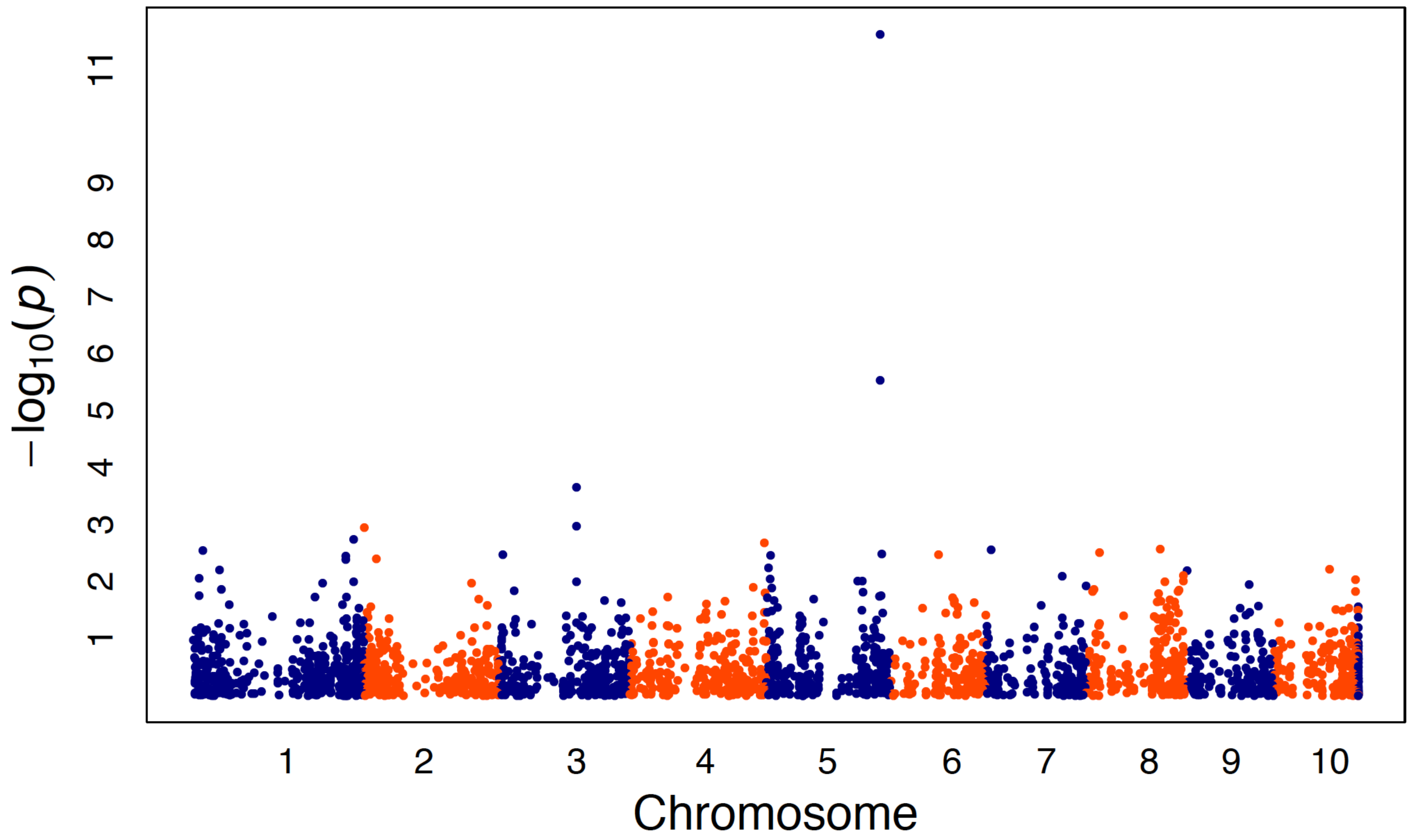
```
Console ~/Desktop/Work/Outreach_Conferences_2016/AfPBA_December_2016/In_class_Example/3.4.2.Tocochromanol_GWAS/Results/
> 
```

- For details on running GAPIT, here is the user manual: http://zzlab.net/GAPIT/gapit_help_document.pdf

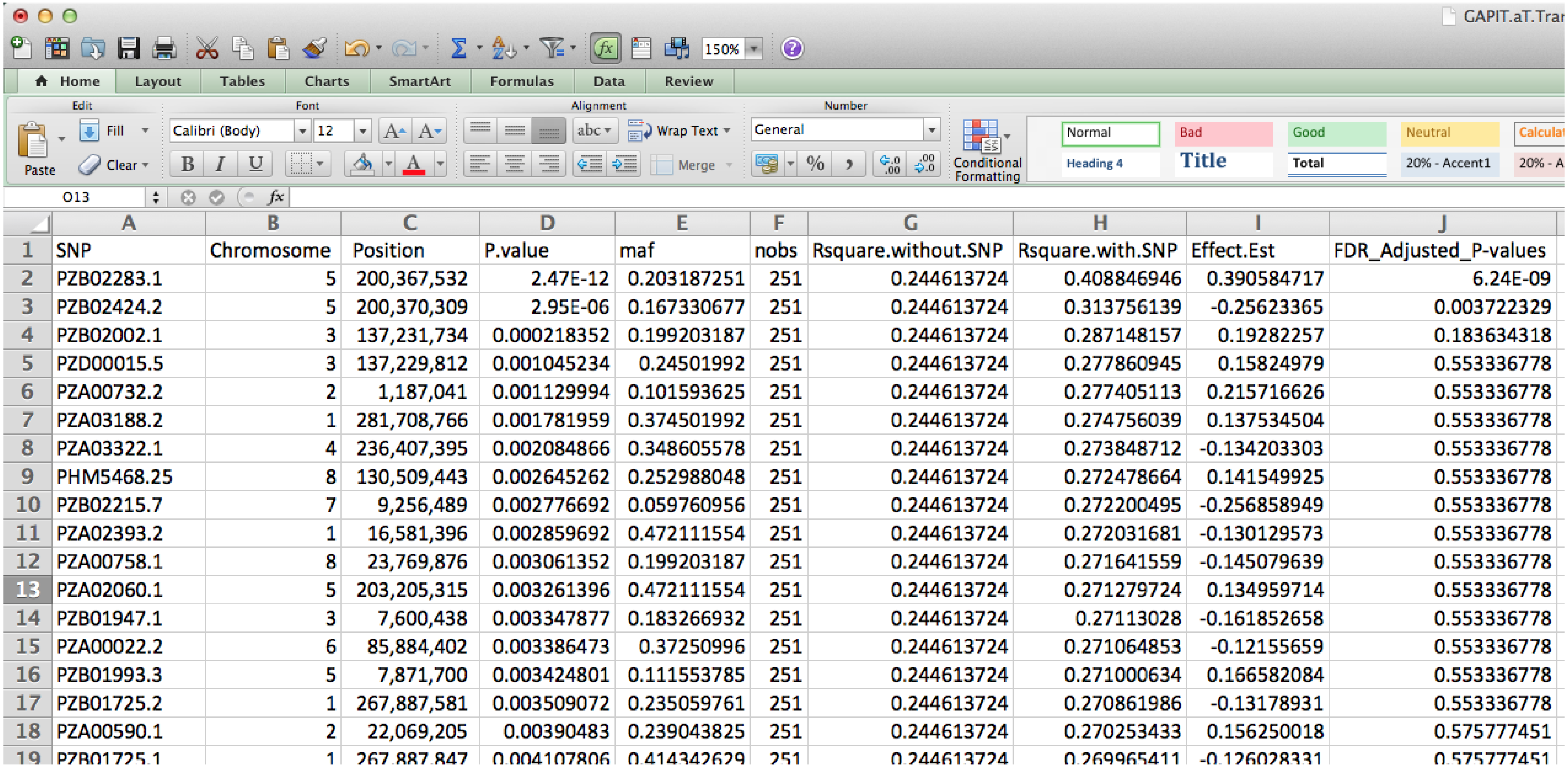
In-class example: GWAS scan of Lipka et al. (2013) data subset



In-class example: GWAS scan of Lipka et al. (2013) data subset

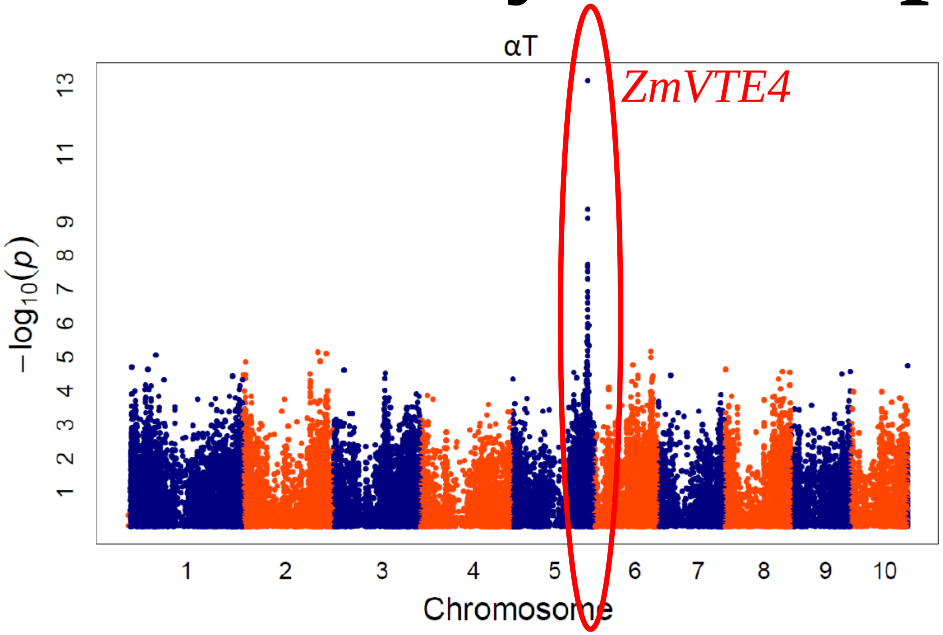


In-class example: GWAS scan of Lipka et al. (2013) data subset



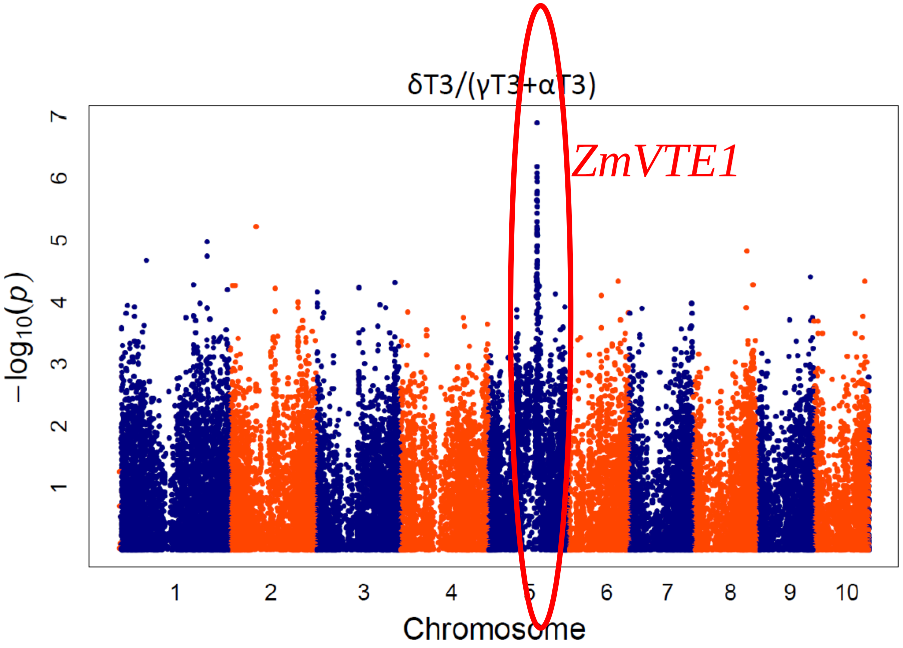
	A	B	C	D	E	F	G	H	I	J
1	SNP	Chromosome	Position	P.value	maf	nobs	Rsquare.without.SNP	Rsquare.with.SNP	Effect.Est	FDR_Adjusted_P-values
2	PZB02283.1	5	200,367,532	2.47E-12	0.203187251	251	0.244613724	0.408846946	0.390584717	6.24E-09
3	PZB02424.2	5	200,370,309	2.95E-06	0.167330677	251	0.244613724	0.313756139	-0.25623365	0.003722329
4	PZB02002.1	3	137,231,734	0.000218352	0.199203187	251	0.244613724	0.287148157	0.19282257	0.183634318
5	PZD00015.5	3	137,229,812	0.001045234	0.24501992	251	0.244613724	0.277860945	0.15824979	0.553336778
6	PZA00732.2	2	1,187,041	0.001129994	0.101593625	251	0.244613724	0.277405113	0.215716626	0.553336778
7	PZA03188.2	1	281,708,766	0.001781959	0.374501992	251	0.244613724	0.274756039	0.137534504	0.553336778
8	PZA03322.1	4	236,407,395	0.002084866	0.348605578	251	0.244613724	0.273848712	-0.134203303	0.553336778
9	PHM5468.25	8	130,509,443	0.002645262	0.252988048	251	0.244613724	0.272478664	0.141549925	0.553336778
10	PZB02215.7	7	9,256,489	0.002776692	0.059760956	251	0.244613724	0.272200495	-0.256858949	0.553336778
11	PZA02393.2	1	16,581,396	0.002859692	0.472111554	251	0.244613724	0.272031681	-0.130129573	0.553336778
12	PZA00758.1	8	23,769,876	0.003061352	0.199203187	251	0.244613724	0.271641559	-0.145079639	0.553336778
13	PZA02060.1	5	203,205,315	0.003261396	0.472111554	251	0.244613724	0.271279724	0.134959714	0.553336778
14	PZB01947.1	3	7,600,438	0.003347877	0.183266932	251	0.244613724	0.27113028	-0.161852658	0.553336778
15	PZA00022.2	6	85,884,402	0.003386473	0.37250996	251	0.244613724	0.271064853	-0.12155659	0.553336778
16	PZB01993.3	5	7,871,700	0.003424801	0.111553785	251	0.244613724	0.271000634	0.166582084	0.553336778
17	PZB01725.2	1	267,887,581	0.003509072	0.235059761	251	0.244613724	0.270861986	-0.13178931	0.553336778
18	PZA00590.1	2	22,069,205	0.00390483	0.239043825	251	0.244613724	0.270253433	0.156250018	0.575777451
19	PZB01725.1	1	267,887,847	0.004107806	0.414342629	251	0.244613724	0.269965411	-0.126028331	0.575777451

GWAS identified signals near two biosynthetic pathway genes



- Peak SNP located 70 bp from *ZmVTE1* start site (P-value = $1.29\text{x}10^{-7}$)
- We are the first to identify *ZmVTE1* in a maize association panel

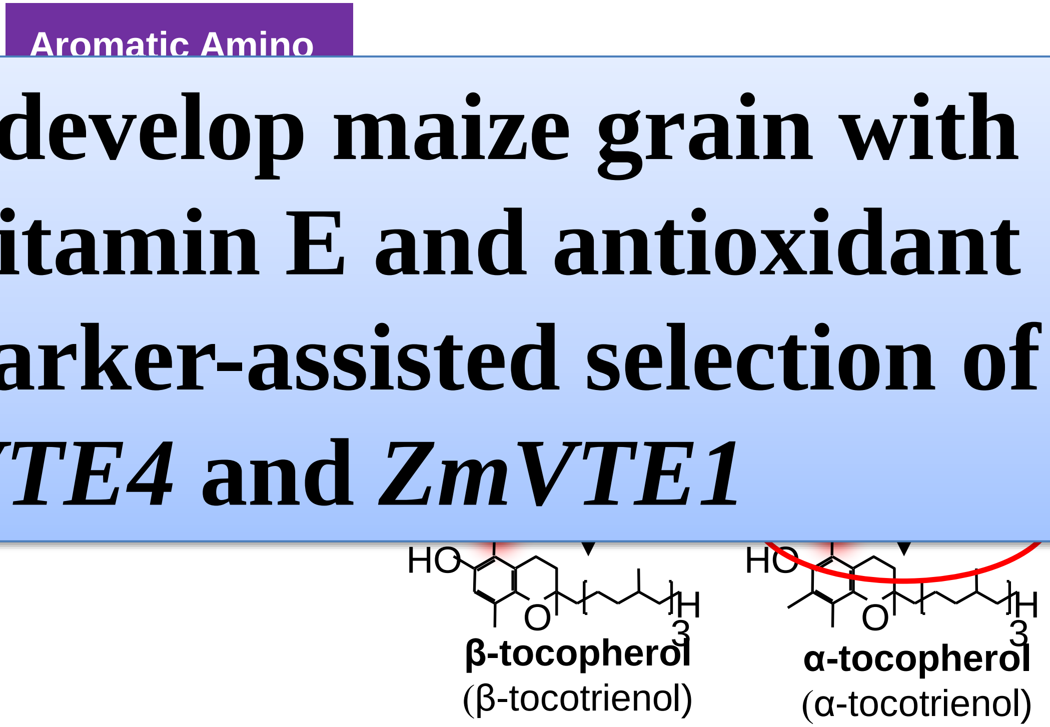
- Peak SNP within *ZmVTE4* (P-value = $7.36\text{x}10^{-14}$)
- *ZmVTE4* has been previously identified



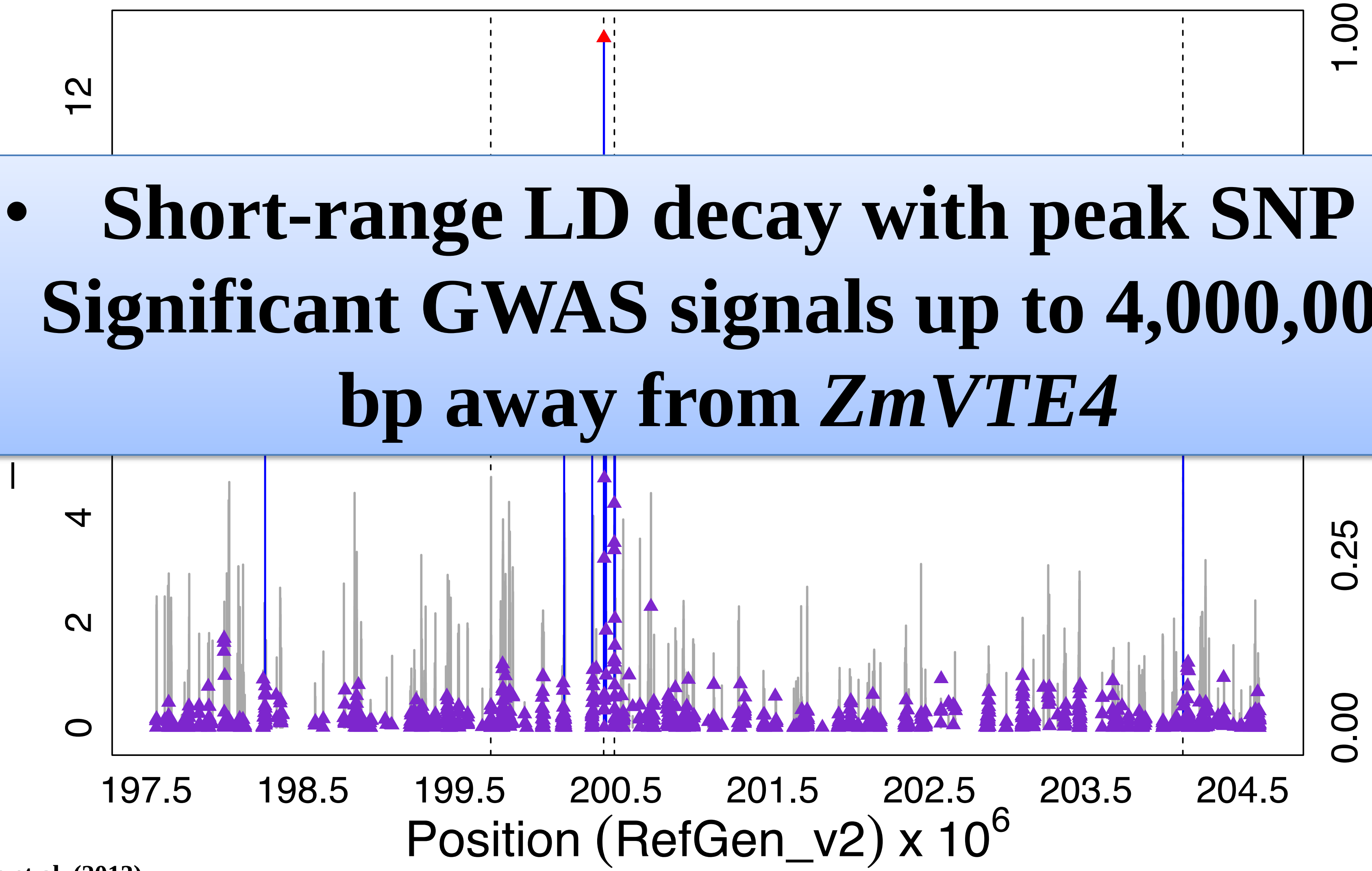
Lipka et al. (2013)

ZmVTE4 and *ZmVTE1* are important genes

- Possible to develop maize grain with enhanced vitamin E and antioxidant levels via marker-assisted selection of *ZmVTE4* and *ZmVTE1*



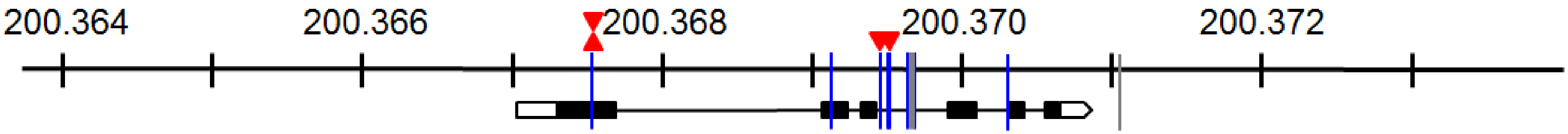
Elucidating the association between αT and *ZmVTE4*



- Short-range LD decay with peak SNP
- Significant GWAS signals up to 4,000,000 bp away from *ZmVTE4*

Lipka et al. (2013)

Stepwise model selection identified two other *ZmVTE4* SNPs associated with αT



γ-tocopherol methyltransferase (*ZmVTE4*)

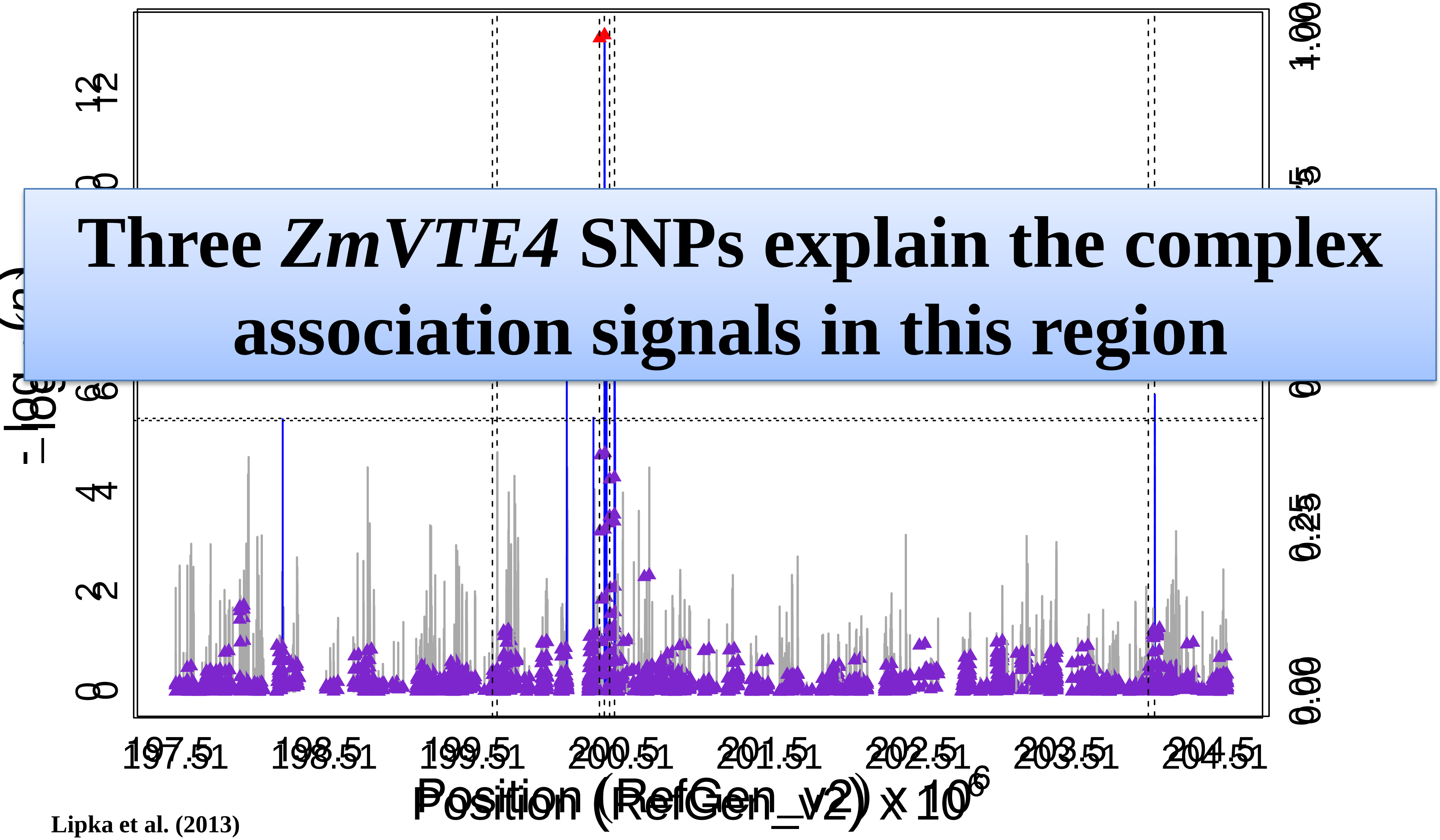
▲ = SNP identified in GWAS

▼ = SNP identified in stepwise model selection (developed in Segura et al., 2012)

- *ZmVTE4* signal explained by three SNPs
- 5.76-fold change in αT levels between most and least favorable haplotypes of these three SNPs

Lipka et al. (2013)

Including three *ZmVTE4* SNPs as covariate removes signal

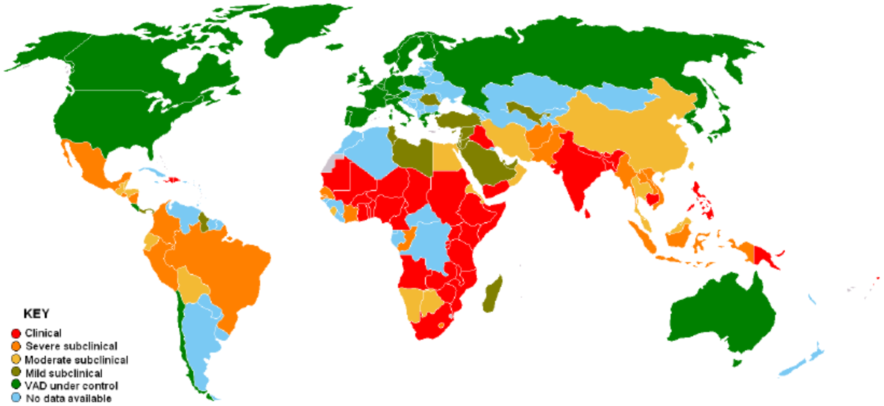


Three *ZmVTE4* SNPs explain the complex association signals in this region

Lipka et al. (2013)

Targeting vitamin A deficiency through biofortification

- Vitamin A deficiency (VAD):
  - Affects 17-30% of children under 5
  - 250-500,000 children become blind every year
  - Infant morbidity and mortality



Source: en.wikipedia.org

- Maize is a primary food source in many vitamin A deficient regions

- Biofortification: breed locally-adapted maize lines for increased provitamin A levels in grain

Work in maize provitamin A biofortification prior to Owens/Lipka et al. (2014)

- Candidate gene studies identified loci in maize (Harjes et al., 2008; Vallabheneni et al., 2010; Yan et al. 2010)

Owens/Lipka et al (2014):

1.) Conduct an GWAS to identify new candidate genes

2.) Determine a minimal marker set to accurately predict carotenoid levels

- Pleiotropy identified among metabolite QTL (Kandianis et al., 2013)

Source: Chandler/Lipka et al., 2013

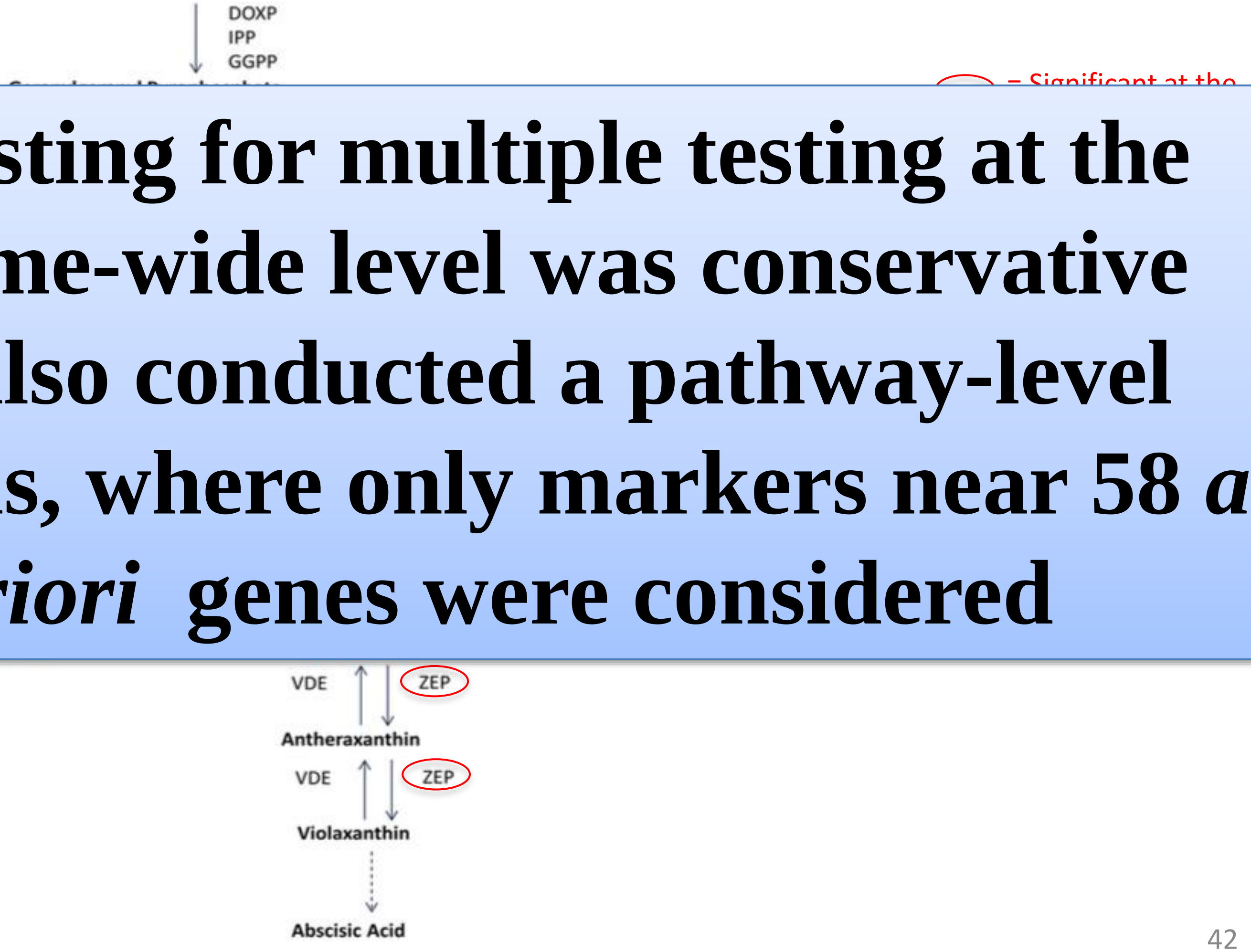
Data analyzed in Owens/Lipka et al. (2014)

- **Maize lines with white kernels do not produce measureable carotenoids**
- **We only analyzed a subset of 201 lines that range from light yellow to dark orange kernel color**

field seasons

- Compound levels quantified in grain:
  - Carotenoids for 252 lines

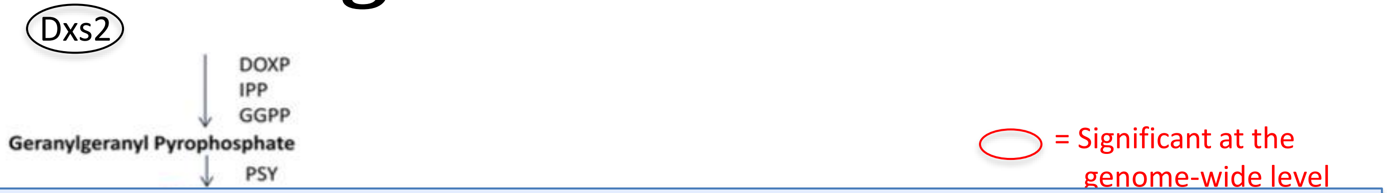
GWAS found significant marker-trait associations near carotenoid pathway genes



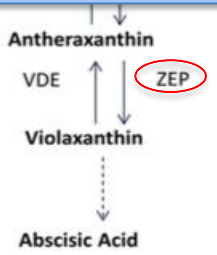
- **Adjusting for multiple testing at the genome-wide level was conservative**
- **We also conducted a pathway-level analysis, where only markers near 58 *a priori* genes were considered**

Owens/Lipka et al. (2014)

GWAS found significant marker-trait associations near carotenoid pathway genes


Dxs2
DOXP
IPP
GGPP
Geranylgeranyl Pyrophosphate
PSY
= Significant at the genome-wide level

- This work identified potential targets for marker-assisted selection (MAS)
- Are selecting for these target loci sufficient for improving provitamin A content in maize grain?



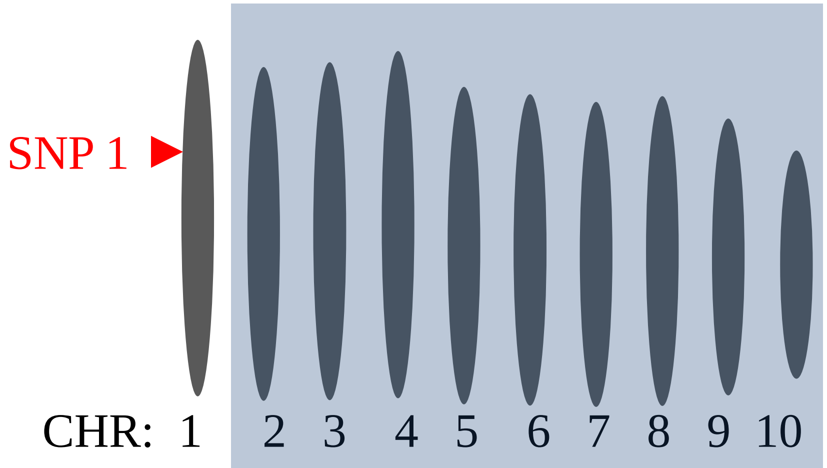
Owens/Lipka et al. (2014)

Targeted marker subsets for estimating kinship

- Suppose we are testing SNP 1 on chromosome 1

- K_chr model has greater power to detect marker-trait associations in high-LD regions

  – Similar "leave one chromosome out" approach used for other chromosomes



Rincent et al. (2014)

Angela Chen

Re-evaluated associations using K_chr model

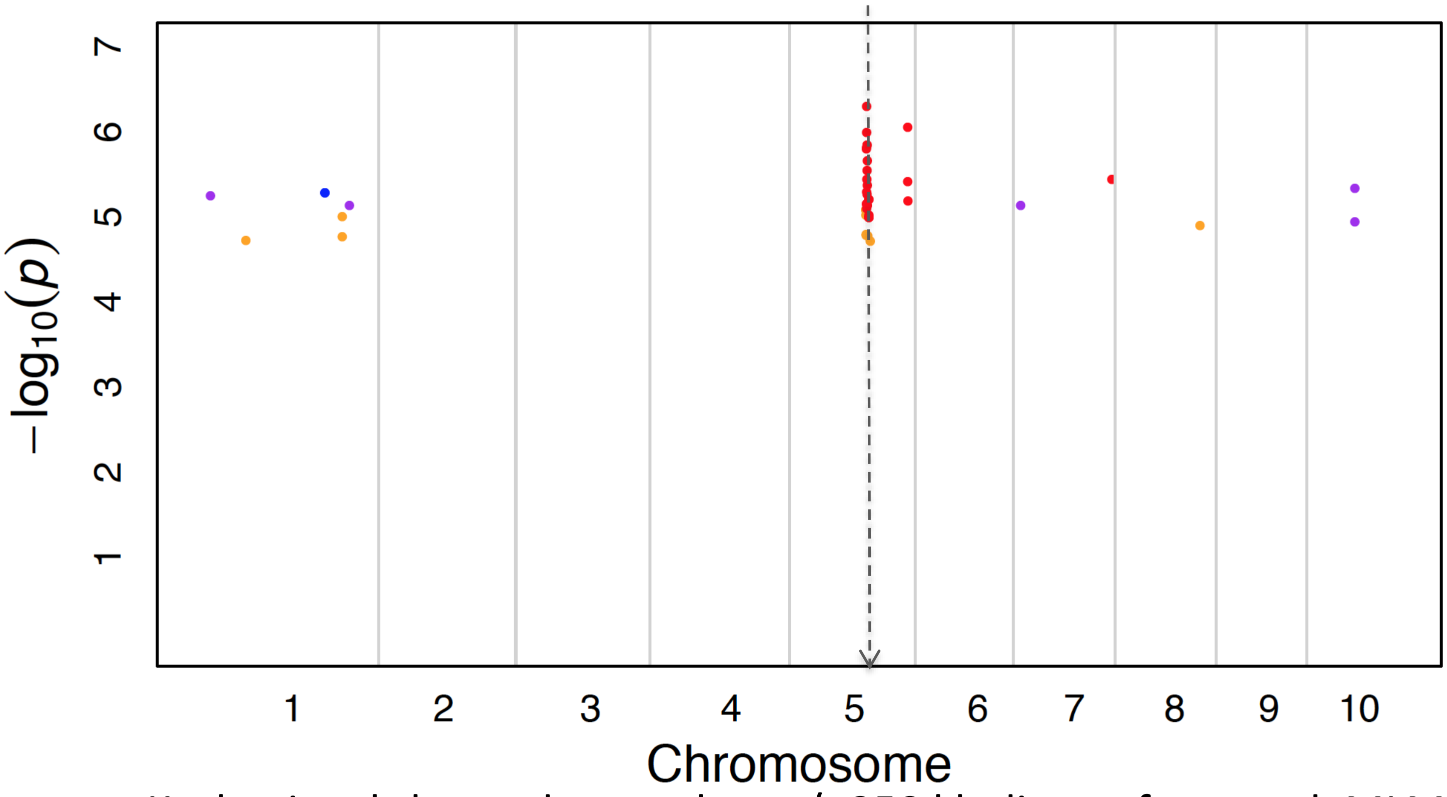
- **Previously published GWAS results from two maize diversity panels:**
  - Mendelian: Sweet vs. starchy corn
  - Polygenic: Carotenoids and tocochromanols
  - Complex: Flowering time and plant height
- **Compared results of the K_chr model to the unified MLM:**
  - Did the K_chr model identify signals in “novel genomic regions”?
  - Did the K_chr model identify more statistically significant associations in high LD regions?

Chen and Lipka (2016)

K_chr identified signals in “novel genomic regions”

Angela Chen

Four tocochromanol traits in Goodman diversity panel



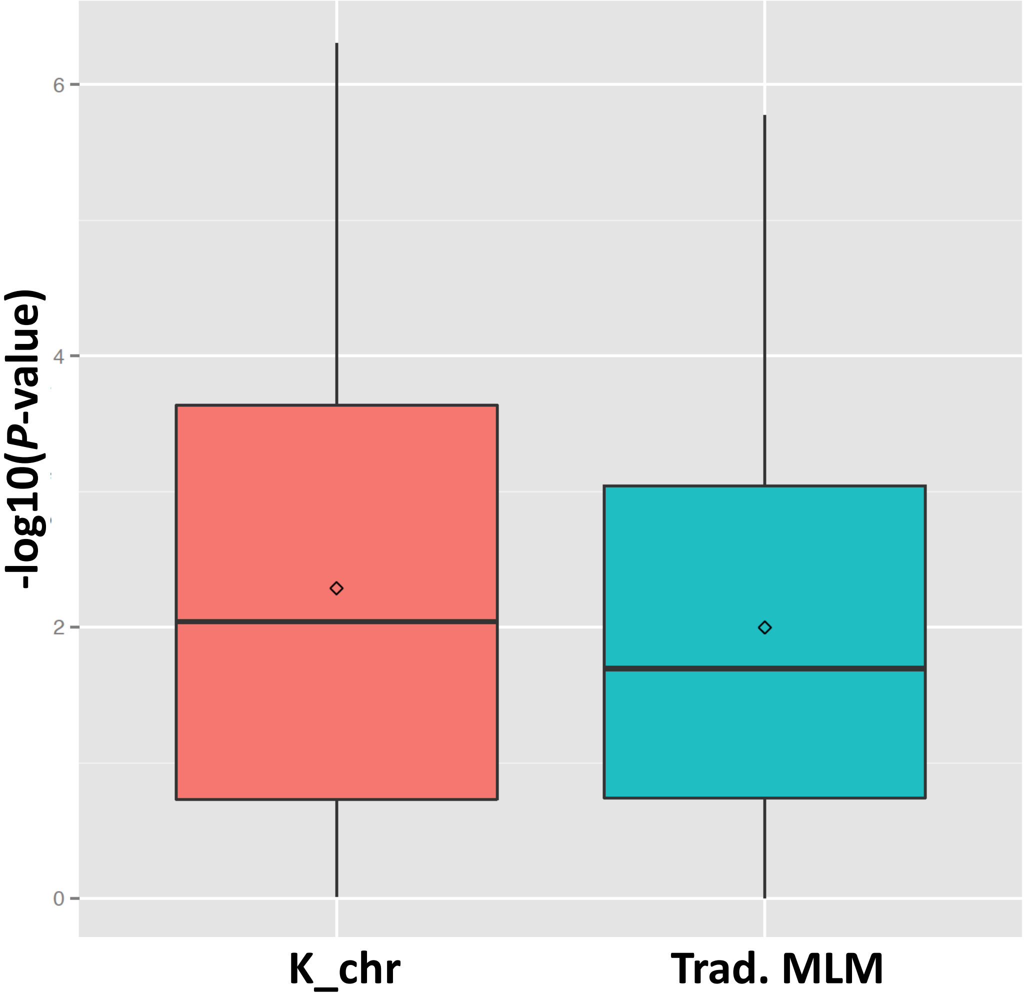
K_chr signals located more than +/- 250 kb distant from trad. MLM signals

Chen and Lipka (2016)

Angela Chen

K_chr identified stronger associations in high LD regions

Associations with tocotrienol ratio in vicinity of *ZmVTE1*



Chen and Lipka (2016)