Method of Analysis of Multi-parent Mapping Populations Affects Detection of QTL

Odell, S. G.[*,1,2], **Praud, S.**[3], **Ross-Ibarra, J.**[2,4,5] **and Runcie, D.**[1]

[1]Dept. of Plant Sciences, University of California, Davis, CA, USA, [2]Dept. of Evolution and Ecology, University of California, Davis, CA, USA, [3]Limagrain, Chappes, France, [4]Center for Population Biology, University of California, Davis, CA, USA, [5]Genome Center, University of California, Davis, CA, USA

ABSTRACT The search for quantitative trait loci (QTL) that explain complex traits such as yield and drought tolerance has been ongoing in all crops. Methods such as bi-parental QTL mapping and genome-wide association studies (GWAS) each have their own advantages and limitations. Multi-parent advanced generation inter-crossing (MAGIC) contain more recombination events and genetic diversity than bi-parental mapping populations and reduce the confounding effect of population structure that is an issue in association mapping populations. Here we discuss the results of using a MAGIC population of doubled haploid (DH) maize lines created from 16 diverse founders to perform QTL mapping, comparing QTL identified using a 600K SNP array to those found using founder probabilities and haplotype probabilities generated by determining the regions of the MAGIC DH lines that were derived from the 16 founders and by identifying regions of identity-by-descent (IBD) between the 16 founders, respectively. The three methods have differing power and resolution for detecting QTL for a variety of agronomic traits. This highlights the importance of considering different approaches to analyzing genotypic datasets, and shows the limitations of binary SNP data for identifying multi-allelic QTL.

KEYWORDS QTL, MAGIC

Introduction

The study of evolutionary quantitative genetics requires the ability to link differences in phenotype to genotypic variation. Natural and artificial selection acts on phenotypes, but only heritable phenotypic variation will result in changes in population means. Maize presents an excellent model organism to answer these questions due to the combination of extensive genetic and phenotypic resources, and the ability to create mapping populations. In addition, maize is one of the most widely produced crops in the world and is a major source of calories for millions of people. Decades of research into maize genetics have resulted in the identification of many quantitative trait loci (QTL) that explain variation in phenotypes such as yield, flowering time, and plant height ((Buckler *et al.* 2009);(Beavis *et al.* 1991);(Martinez *et al.* 2016)). Such traits are extremely agronomically important. They are also crucial plant phenotypes in terms of fitness and local adaptation. Researchers have been able to discover large-effect QTL for a number of agronomic traits through the use of different types of mapping populations (Wallace *et al.* 2014). The choice of any population comes with associated advantages and limitations. In particular, they tend to vary in two main characteristics: (1) their ability to capture genetic diversity and (2) their power to detect QTL of small effect. Multi-parent Advanced Generation Intercross (MAGIC) populations have been used in breeding to increase the genetic diversity and number of recombination events included in a mapping population compared to biparental populations (Huang *et al.* 2012);(Huang *et al.* 2012);(Aylor *et al.* 2011);(Dell'Acqua *et al.* 2015);(Highfill *et al.* 2016). Compared to genome-wide association panels, MAGIC populations have more power to detect rare alleles (i.e. alleles that are only present in one of the parents) and can better estimate allelic effects because the crossing scheme increases the frequency of all parental alleles to be approximately equal. Simulations of an 8-parent MAGIC population showed that sample sizes of 300 could detect QTL accounting for 12% of variance with a power of 82% (Dell'Acqua *et al.* 2015). Lastly, a MAGIC population avoids confounding due to population structure that is encountered with GWAS because the pedigree of the lines is known. In this study, we utilized a MAGIC population of about 400 doubled-haploid lines derived from 16 inbred maize parents developed by Limagrain to understand how methods of representing genotypic data can impact the identification of QTL. Extensive genetic resources already exist for maize, but do not possess the same diversity and statistical power as the Biogemma MAGIC population. A maize nested association mapping (NAM) population exists, consisting of RIL populations derived from 25 inbred parents crossed to B73 (Yu *et al.* 2006). Only two inbred parents overlap between the NAM and Biogemma MAGIC populations (B73 and Oh43), and compared to the NAM, the MAGIC population can have similar power to the NAM using half the number of samples (Dell'Acqua *et al.* 2015). Likewise, another maize MAGIC population has previously been created, which overlaps by three parents (A632, B73, and B96) (Dell'Acqua *et al.* 2015). However, the previous MAGIC is derived from 8 inbred maize parents instead of 16, and consists of RILs, not doubled haploids, so some residual heterozygosity may exist. For these reasons, the Biogemma MAGIC population has great potential to reveal new insights into the

* Dept. of Plant Sciences and Dept. of Evolution and Ecology, University of California, Davis, CA, USA E-mail: sgodell@ucdavis.edu

genetic control of quantitative traits in maize. It also serves a reliable standard for QTL mapping method comparision because DH lines were used and the crossing scheme ensured that there is a reasonably even distribution of the 16 founders within the population (supplemental figure?)

Description of the flowering time QTL, *vgt1* and previous research... Using a well-characterized flowering time QTL with a strong candidate causal variant that is variable in the population, we demonstrate differences between the three methods and explore potential epistatic interactions between *vgt1* and other genetic variation in the population.

Materials and Methods

Genotype Data

The MAGIC population was derived from 16 inbred maize parents representing the diversity of European flint and U.S. dent heterotic groups. The 16 founder lines were crossed in a funnel crossing scheme, and then the resulting synthetic population was intercrossed for 3 generations with around 2000 individuals per cycle (Figure 1A). Finally, 800 lines were selected from the synthetic population to create doubled haploids (DH), resulting in 550 MAGIC DH lines at the end of the process. The 16 founder lines and the MAGIC DH lines were all genotyped with the Affymetrix 600K Axiom SNP array (Unterseer *et al.* 2014). In addition, the 16 founder lines were sequenced with Illumina short-read sequencing to a depth of [?]x, resulting in 45.4 millions SNPs and 5.4 indels after filtering using GATK best practices.

Phenotype Data

The MAGIC DH lines were crossed to a tester MBS84 to produce 325 hybrids. Due to variation in flowering time, a subset of the lines could not be crossed to the tester. The hybrids were grown in five different field locations in four different years, resulting in six distinct environment-years. The environment-years included Blois, France in 2014 and 2017, Nerac, France in 2016, St. Paul, France in 2017, Szeged, Hungary in 2017, and Graneros, Chile in 2015. For each genotype, two blocks of around 80 plants were grown under well-watered conditions. Measured phenotypes included days to anthesis (DTA), days to silking (DTS), plant height, percent harvest grain moisture, grain yield, and thousand kernel weight (adjusted to 15% humidity), where values were averaged over blocks. Both flowering time phenotypes were measured as the sum of degree days since sowing with a base temperature of 6°C (48°F). Days to anthesis was considered as the growing degree days until 50% of plants in a block were flowering at 25% of the central tassel spike.

Calculation and Validation of Founder Probabilities

The package R/qtl2 (Broman *et al.* 2019) was used to determine founder probabilities of the DH lines using the 600K genotype data and the cross type "riself16". Due to the fact that the actual crossing scheme and the cross type input into R/qtl2 differed, we wanted to assess the accuracy of the founder probabilities. This was done by simulating lines using the actual crossing scheme and assessing the performance of the calc_genoprobs function of R/qtl2 in correctly identifying the founder genotype (Figure 1). We developed an R package, magicsim (https://github.com/sarahodell/magicsim), to simulate the lines using the maize consensus genetic map from (Ogut *et al.* 2015) to generate approximate recombination rates across the chromosome. For 400 simulated lines, 99.6% of SNPs were correctly assigned to the founder genotype. The founder probabilities were filtered for LD using an iterative approach where a SNP was dropped if the correlation of probabilities between it and the kept SNP was greater than 0.95. After filtering, a total of 2,929 sites were kept to represent segments of chromosomes with little recombination in the MAGIC DH lines.

Calculation of Identity-by-Descent and Haplotype Probabilities

The use of founder probabilities makes the assumption that all 16 founders have distinct haplotypes. This assumption is not realistic, especially considering the known varying degrees of relatedness between the founders. The identification of regions of shared genetic sequence between founder pairs would allow for the collapsing of founder probabilities into haplotype probabilities. These haplotype probabilities have the potential to increase statistical power by reducing the number of tests performed in QTL mapping. Areas of uncertainty in founder probabilities of the DH lines were associated with regions of identity-by-descent (IBD) between two or more founder lines in that region of the chromosome. IBD was measured from the WGS data of the founders using the software RefinedIBD (Browning and Browning 2013). RefinedIBD requires no missing data, so the software PLINK was used to impute missing data without a reference, which sets all missing sites to the allele frequency of the alterante allele within the samples. In addition, only bi-allelic sites were kept for the calculation of IBD, resulting in a total of 16.9 million SNPs for comparison of sequence similarity. The resulting segments of pairwise IBD between each of the 16 founders were used to identify distinct haplotype blocks [elaborate, supplemental figure?]. Within blocks, the founder probabilities for founders that were in IBD were summed to obtain haplotype probabilities. This resulted in haplotype blocks with the number of unique haplotypes within blocks ranging from 6 at the lowest and 16 at the highest. The haplotype probabilities were filtered for LD using an iterative approach where, for all haplotype blocks with the same number of distince haplotypes,a SNP was dropped if the correlation of probabilities between it and the kept SNP was greater than 0.95. After filtering, a total of 9,368 sites were kept to represent haplotype blocks in the MAGIC DH lines.

Imputation of WGS Data

The founder probabilities obtained from the 600K genotype data of the founders and the MAGIC DH lines was used to impute the WGS data from the founders onto the 325 MAGIC lines. This was done using a custom script which calculated the probabilities of a MAGIC line having the alternate allele at each site in the WGS data based upon the founder probabilities of the line and the alleles of the 16 founders at that site. Only sites with less than 25%(?) missing data were kept for imputation.

Association and QTL Mapping

The R package GridLMM was used to run association mapping using the three different methods of representing the genotype data (Runcie and Crawford 2019). The function GridLMM_ML was used with the "ML" option. The model for each was

$$Y = X\beta + Zu + \epsilon \quad (1)$$

where Y is the response variable, X is the genotype matrix, β, is the effect size, Z is the design matrix, u is the random effects, and ϵ is the error. Significance cutoffs for p-values were obtained using permutation testing, taking the 5% cutoff from 1000 randomized permutations for each method. Code for the analyses

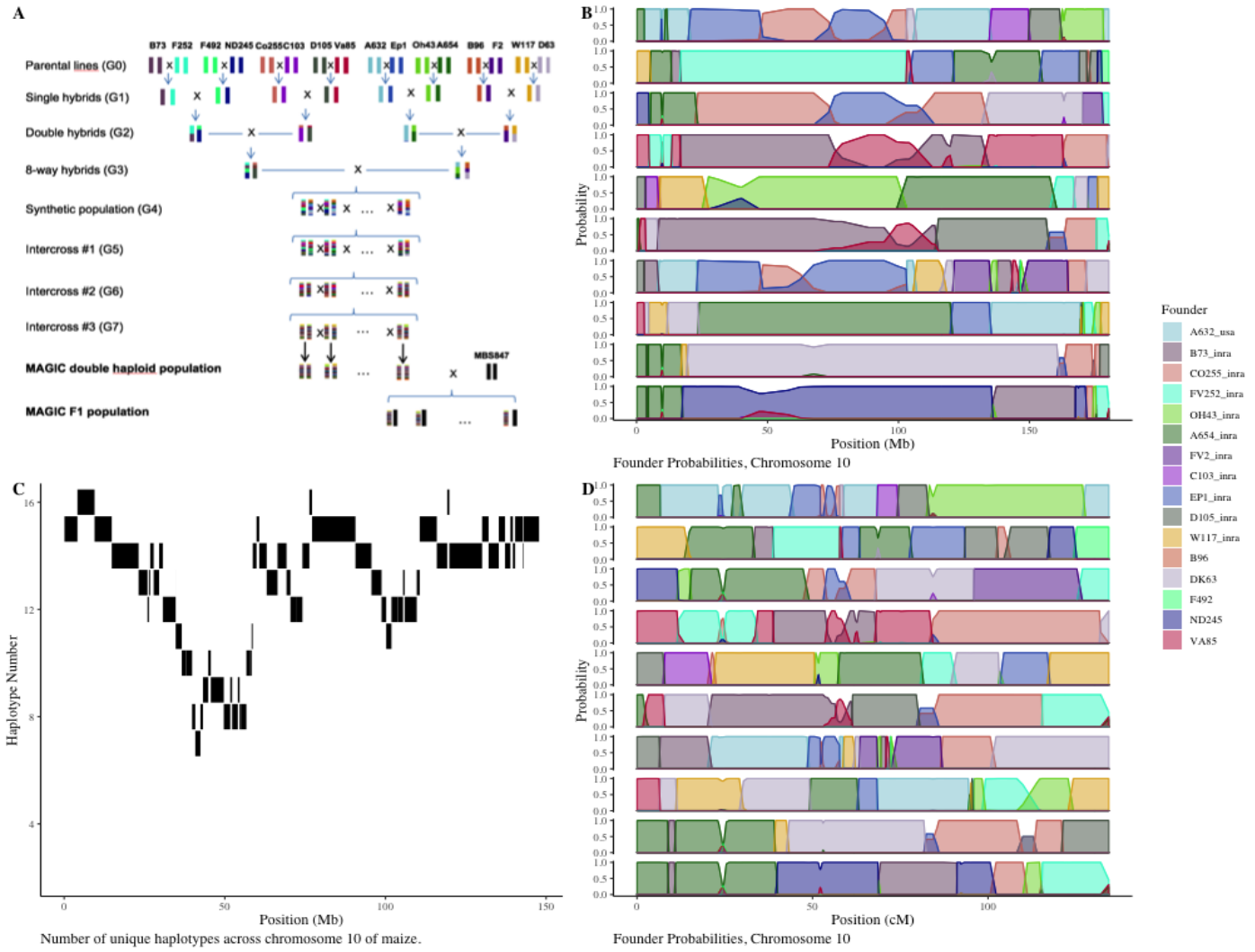

Figure 1 Structure, diversity, and founder representation of the MAGIC population(**A**) The crossing scheme of the MAGIC population. (**B**) Founder probabilities for 10 MAGIC DH lines on chromosome 10 in physical distance. (**C**) The number of unique haplotype groups among the 16 founder lines across chromosome 10. (**D**) Founder probabilities for 10 MAGIC DH lines on chromosome 10 in genetic distance.

can be found [github link]. We used the software GEMMA [reference] to build a mixed linear model to identify QTL due the faster run time of this software with larger datasets.

Method Comparison

The results of the three methods of identifying QTL were compared using two main criteria: (i) presence or absence of identified QTL peaks and (ii) the resolution of those QTL peaks. QTL bounds were determined by identifying the most significant SNP for a QTL peak and demarcating the left and right bounds of the QTL as the left-most and right-most SNPs within a 100Mb window centered on the highest SNP that have a log10 p-value that is 2 log10 p-values below that of the highest SNP.

The presence and absence of QTL was compared across the three methods for each phenotype. A QTL was said to be identified across methods if the QTL bound for that QTL overlapped between the three methods.

Results

MAGIC Population

Analysis of the MAGIC population showed that the representation of the 16 founders in the MAGIC DH lines was relatively even, with the highest percentage founder, A654, representing 6.741% and the lowest percentage founder, EP1, representing 5.237% (Supplemental Figure). Across all regions of the each chromosomes, for 99%(?) of markers in the 600K array, at least one DH line was derived from each of the 16 founders with high confidence (greater than 0.9 probability) (Supplemental Figure). One notable exception was on chromosome 6, where a region [location] had no representation from [founders]. This may be due to these founders being in close IBD with another founder, resulting in uncertainty in the founder probablities. Another possible explanation is that there was selection against theese founders during the breeding process. This region approximately corresponds to the NOR region [citation].

The founder probabilities determined using R/qtl2 were able to assign founders to the DH lines with high confidence (>0.90)

for [percentage] of the 10 chromosomes of maize.

Simulation and Validation of Founder Probabilities

Using the package we created, magicsim, we simulated chromosome 10 for 1000 MAGIC DH lines following the crossing scheme used for the actual population. We then the ran R/qtl2 on these simulated lines to obtain founder probabilities and compared these probabilities to the known founder identities of the simulated lines. In total, for sites for which R/qtl2 was able to assign founder probabilities, the maximum probability founder matched the actual founder for 99.8% of sites. This reinforced our confidence in the founder probabilties obtained from the actual data.

Identity-By-Descent and MAGIC Haplotypes

The results showed that a total of [bp][cM] of genome were in IBD between at least two different founders. The average size of an IBD segment between two founders was 140kb (0.51 cM) with a median of 122kb (0.46). Pairwise IBD segments sizes ranged from 8 kb (0.3 cM) to 673 kb (1.61 cM). The total percentage of IBD between two founders ranged from 0.0018% (F492 and VA85) to 4.3858% (B73 and A632), with an average of 0.06130%. There were no IBD segments found for 18 of 120 possible pairwise founder combinations. [Distribution across chromosomes]

Amount of IBD between the founders and the tester, MBS847...

Phenotype Data

The distribution of the six measured phenotypes were approximately normal (Figure 2).

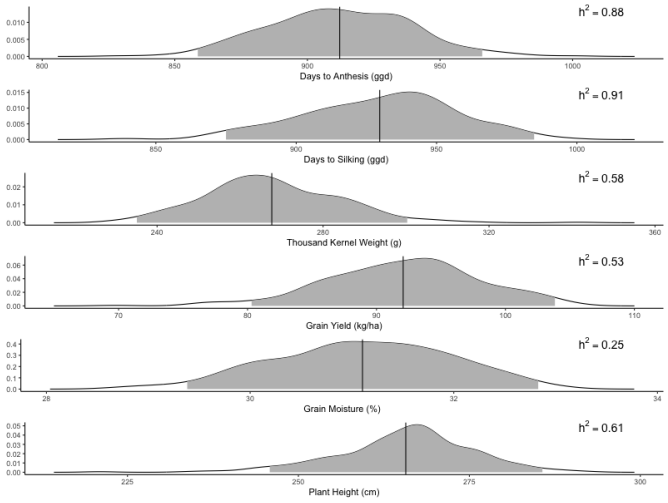

Figure 2 Distribution of phenotypes The density plots of the six measured phenotypes. The vertical bar represents the mean, and the grey shading shows two standard deviations. The heritability of each trait is shown on the right.

QTL Mapping and Association Mapping

The three methods varied both in their ability to identify QTL and the resolution of those QTL.

Resolution of QTL bounds differs between methods. The GWAS methods should be most powerful at identifying QTL for which the causal variant is biallelic and the tagged SNP is in tight LD with the causal varaint. However, for multi-allelic QTL or QTL for which LD is low between tagged SNPs, this method begins to lose power. Founder probabilities increase the odds of detecting both QTL that are multi-allelic and QTL whose causal variant is not in tight LD with any one tagged SNP [Supplemental Figure?]. Lastly, haplotype probabilities potentially improve on the power of founder probabilities to detect QTL that meet the above criteria by reducing the number of test. However, haplotype probabilities also may obscure the signal of some QTL if founders are called as being in IBD with one another when they actually differ for the causal variant. Due to the fact that the founder and haplotype probabilities take into account recent recombination events, whereas the GWAS method only uses historical recombination, we predicted that founder and haplotype mapping would result in higher resolution around QTL peaks. Higher resolution QTL are ideal in that it makes it easier to narrow down candidate genes and potential causal variants when the significant window is smaller.

Variation around vgt1

A previously characterized QTL, *vgt1* is associated with variation in flowering time [citations]. An earlier flowering phenotype is strongly correlated with a MITE insertion about 70kb upstream of the flowering time regulator, *ZmRAP2.7*, an *APETALA*-like transcription factor [].

Evidence of epistasis around *vgt1*. Imperfect association between the presence of the MITE and the predicted effect size of the founder at vgt1 (Figure 4)

Expression of *ZmRAP2.7*

Discussion

Some more test

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

another figure

This is a full size figure (Figure S1) just add stars to the figure

Acknowledgments

We acknowledge the support of our coffee maker that made this work possible

References

Aylor, D. L., W. Valdar, W. Foulds-Mathes, R. J. Buus, R. A. Verdugo, *et al.*, 2011 Genetic analysis of complex traits in the emerging collaborative cross. Genome Research **21**: 1213–1222.

Beavis, W. D., D. Grant, M. Albertsen, and R. Fincher, 1991 Quantitative trait loci for plant height in four maize populations and their associations with qualitative genetic loci. Theoretical and Applied Genetics **83**: 141–145.

Broman, K. W., D. M. Gatti, P. Simecek, N. A. Furlotte, P. Prins, *et al.*, 2019 R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multiparent populations. Genetics **211**: 495–502.

Browning, B. L. and S. R. Browning, 2013 Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics **194**: 459–471.

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.

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.

Highfill, C. A., G. A. Reeves, and S. J. Macdonald, 2016 Genetic analysis of variation in lifespan using a multiparental advanced intercross drosophila mapping population. BMC Genetics **17**: 113.

Huang, B. E., A. W. George, K. L. Forrest, A. Kilian, M. J. Hayden, *et al.*, 2012 A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnology Journal **10**: 826–839.

Martinez, A. K., J. M. Soriano, R. Tuberosa, R. Koumproglou, T. Jahrmann, *et al.*, 2016 Yield qtlome distribution correlates with gene density in maize. Plant Science **242**: 300–309.

Ogut, F., Y. Bian, P. J. Bradbury, and J. B. Holland, 2015 Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. Heredity **114**: 552–563.

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

Unterseer, S., E. Bauer, G. Haberer, M. Seidel, C. Knaak, *et al.*, 2014 A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k snp genotyping array. BMC Genomics **15**: 823.

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.

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics **38**: 203–208.

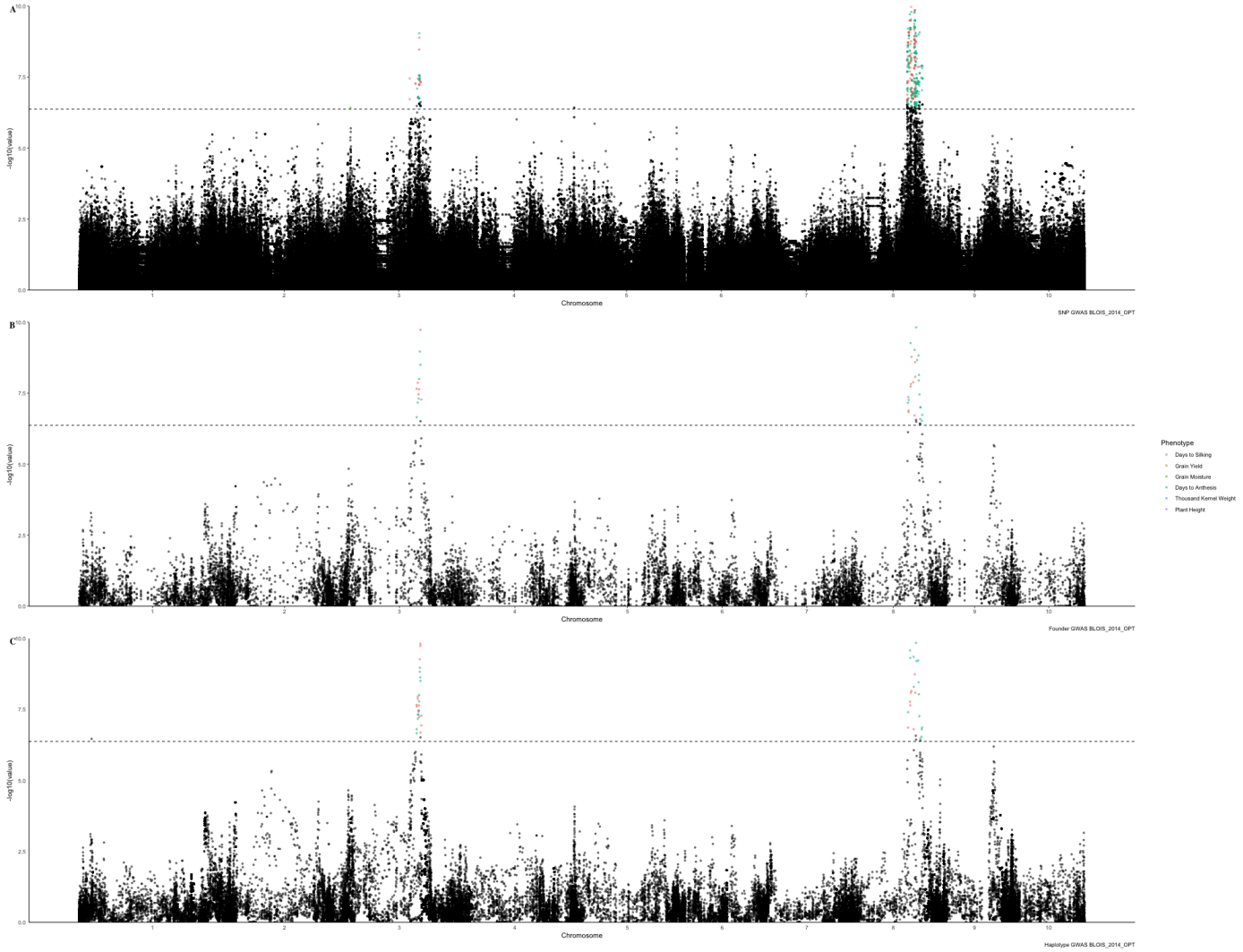

Figure 3 Results of three methods of QTL identification Colored points represent signficant SNPs above the 5% signficance threshold from 1000 random permutations **top** GWAS results using the 600K SNP array. **middle** Results from QTL mapping using founder probabilities **bottom** Results from QTL mapping using haplotype probabilities

Supplement

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

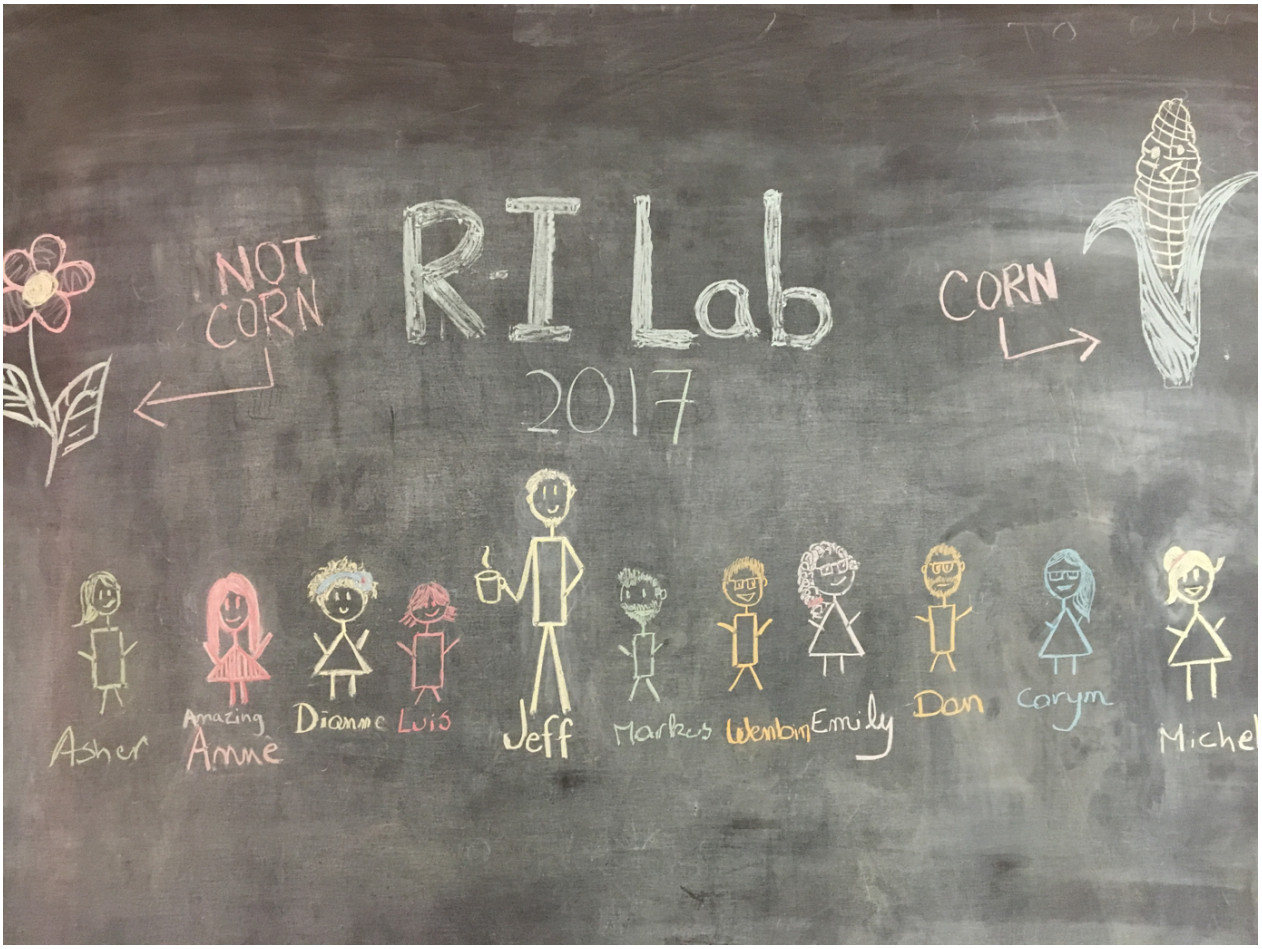

Figure S1 Supplemental figure Test test test

Table S1 Shrink a large table to fit the page

Parameter	Description
Adaptation	**Trait related parameters**
Time to optimum	Generations until new optimum is reached
Adaptation rate (haldane)	Adaptation rate until new optimum is reached. Calculated as $rate(h) = \frac{\frac{ln(x_2)}{sd_{x_{12}}} - \frac{ln(x_1)}{sd_{x_{12}}}}{t_2 - t_1}$
Final genetic variance	Genetic variance in the final generation
Fixations	**Mutations that fix after the optimum shift**
From new mutations (#)	Sum of fixed mutations in the final population that were already segregating before the optimum shift
From standing variation (#)	Sum of fixed mutations in the final population that arose after the optimum shift
Max. effect size	Maximal effect size of all fixations
Mean effect size	Mean effect size of all fixations
Mean effect size of negative fixations	Mean effect size of negative mutations
Mean effect size of positive fixations	Mean effect size of positive mutations
Mean emergence time	Mean generation when a mutation arose that fixed in the last 0.1 N generations
Mean fixation time	Mean generation in which a mutation fixed
Min. effect size	Minimal effect size of all fixations
Negative (#)	Sum of fixed mutations with negative effects in the final population
New/standing fixations	Ratio of mutations from new mutations vs. standing mutations
Proportion negative	Proportion of negative fixations from all mutations
Positive (#)	Sum of fixed mutations with positive effects in the final population
SD of effect sizes	Standard deviation of effect sizes of all fixations
SD of negative effect sizes	Standard deviation of effect sizes of negative fixations
SD of positive effect sizes	Standard deviation of effect sizes of positive fixations
Total (#)	Sum of fixed mutations in the final population
Sweeps	**Mutations that fix faster than 99% of neutral fixations**
Hard sweeps (#)	Sum of selective sweeps from new mutations
Proportion of hard sweeps	Porportion of hard selective sweeps of all selective sweeps
Proportion of sweeps from standing	Proportion of selective sweeps from stainding variation of all selection sweeps
Sweeps (#)	Sum of selective sweeps
Sweeps from standing variation (#)	Sum of selective sweeps from mutations that were already segregating before the optimum shift
Sweeps/fixations	Ratio of sweeps vs. fixations
Segregating sites	**Mutations that segregate in the final generation**
Max. effect size	Maximal effect size of segregating sites
Mean effect size	Mean effect size of segregating sites
Mean effect size of negative sites	Mean effect size of segregating sites with negative effects
Mean effect size of positive sites	Mean effect size of segregating sites with positive effects
Mean frequency of all sites	Mean allele frequency of segregating sites
Mean frequency of negative sites	Mean allele frequency of segregating sites with negative effects
Mean frequency of positive sites	Mean allele frequency of segregating sites with positive effects
Min. effect size	Minimal effect size of segregating sites
Negative (#)	Sum of segregating sites with negative effect
Positive (#)	Sum of segregating sites with positive effect
Proportion of negative sites	Proportion of segregating sites with negative effect of all segregating sites
Standard deviation of effect sizes	Standard deviation of effect sizes of all segregating sites
Total (#)	Sum segregating sites in the final generation