Genome Wide Association Studies Using a New Nonparametric Model Reveal the Genetic Architecture of 17 Agronomic Traits in an Enlarged Maize Association Panel

Ning Yang1*, Yanli Lu1,2*, Xiaohong Yang3, Juan Huang1, Yang Zhou1, Farhan Ali1, Weiwei Wen1, Jie Liu1, Jiansheng Li3, Jianbing Yan1*

1 National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China, 2 Maize Research Institute, Sichuan Agricultural University, Wenjiang, Sichuan, China, 3 National Maize Improvement Center of China, Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing, China

Abstract

Association mapping is a powerful approach for dissecting the genetic architecture of complex quantitative traits using high-density SNP markers in maize. Here, we expanded our association panel size from 368 to 513 inbred lines with 0.5 million high quality SNPs using a two-step data-imputation method which combines identity by descent (IBD) based projection and k-nearest neighbor (KNN) algorithm. Genome-wide association studies (GWAS) were carried out for 17 agronomic traits with a panel of 513 inbred lines applying both mixed linear model (MLM) and a new method, the Anderson-Darling (A-D) test. Ten loci for five traits were identified using the MLM method at the Bonferroni-corrected threshold $-\log_{10}(P) > 5.74$ ($\alpha = 1$). Many loci ranging from one to 34 loci (107 loci for plant height) were identified for 17 traits using the A-D test at the Bonferroni-corrected threshold $-\log_{10}(P) > 7.05$ ($\alpha = 0.05$) using 556809 SNPs. Many known loci and new candidate loci were only observed by the A-D test, a few of which were also detected in independent linkage analysis. This study indicates that combining IBD based projection and KNN algorithm is an efficient imputation method for inferring large missing genotype segments. In addition, we showed that the A-D test is a useful complement for GWAS analysis of complex quantitative traits. Especially for traits with abnormal phenotype distribution, controlled by moderate effect loci or rare variations, the A-D test balances false positives and statistical power. The candidate SNPs and associated genes also provide a rich resource for maize genetics and breeding.

Introduction

Maize (Zea mays L.) is one of the most important food, feed and industrial crops globally. Grown extensively under different climate conditions across the world, maize shows an astonishing amount of phenotypic diversity [1]. Identifying the underlying natural allelic variations for the phenotypic diversity will have immense practical implications in maize molecular breeding for improving nutritional quality, yield potential, and stress tolerance.

With the rapid development of next generation sequencing and high-density marker genotyping techniques, there emerges tremendous interest in using association mapping to identify genes responsible for quantitative variation of complex traits [2]. The use of GWAS has been well demonstrated in model plants such as Arabidopsis [3] and rice [4]. In maize, we examined the genetic architecture of maize oil biosynthesis in 368 diverse maize inbred lines with over 1.06 million SNPs obtained from RNA sequencing and DNA array using the GWAS strategy [5]. Despite the great potential that GWAS has to pinpoint genetic polymorphisms underlying agriculturally important traits, false discoveries are a major concern and can be partially attributed to spurious associations caused by population structure and unequal relatedness among individuals in a given panel [6]. A number of statistical approaches have been proposed, among which the mixed linear model (MLM) is one of the popular methods that can eliminate the excess of low p values for most traits [6,7]. However, Zhao et al. [8] performed GWAS using a NAIVE model in each subpopulation and MLM with inferred population structure as a fixed effect in the whole mapping panel of rice, and their results suggested that MLM may lead to false negatives by overcompensating for population structure and relatedness. To improve the MLM, some strategies to best utilize marker data have been...
Phenotypic variation for 17 agronomic traits

The objectives of this research were (1) to explore an efficient SNP projection and imputation method to inference missing genotypes for the 145 inbreds based on modest density markers. However, detailed instructions for the algorithm were not published. Most current GWAS methods lack the power to detect rare alleles and this has limited the application of GWAS, since rare alleles are common in maize diversity collections [1,5]. Parametric tests of association are sensitive to SNPs with minor allele frequencies, which can artificially increase association scores. Balancing samples across population subdivisions can homogenize allele frequencies, elevating rates of globally rare variants that are common in certain subdivisions [5].

In this study, 513 diverse maize inbred lines [13], representing tropical/subtropical and temperate germplasm, were genotyped by MaizeSNP50 BeadChip containing 56,110 SNPs [14]. RNA sequencing (RNA-seq) was performed on 368 out of the 513 maize inbreds using 90-bp paired-end Illumina sequencing, resulting in 2,445.9 Gb of raw sequencing data. 536,809 high quality SNPs obtained by combining the two genotyping platforms (RNA-seq and SNP array) [5,16] were used in the study. For the additional 145 maize lines, the genotype calls of unique loci from the integrated SNP data were projected based on regions of IBDFD to physical maps constructed using 56110 SNPs, and then high-density markers with more than 0.5 million SNPs were obtained for all the lines. Out of 56,110 SNPs from MaizeSNP50 data set, 49728 SNPs overlapped with the integrated SNPs data based on their physical positions (B73 RefGen_v2). The 49,728 common SNPs were regarded as core or frame markers for projection based on IBDFD regions. In order to evaluate the performance of IBDFD [17] based projection, training and validation datasets were established for chromosome 1, which had 7618 core markers from Illumina Maize SNP50 and 88581 SNPs from the integrated data set. The genotypes for one maize line with RNA-seq data in IBDFD regions were assigned to the matched target line without RNA-seq data for each SNP. The projection accuracy was calculated by comparing inferred genotypes of 368 lines with their real genotype obtained from RNA-seq. In addition, KNN algorithm [4] which infers a large number of missing genotypes generated from low-coverage genome sequencing was used to impute the missing genotypes of the unique loci from RNA-seq SNP data based on 49728 frame markers. Single method analysis, either IBDFD based projection or KNN algorithm, cannot achieve both optimal accuracy and coverage (see Materials and Methods). However, the combination of IBDFD based projection and KNN seemed effective to infer a large number of missing genotypes. In order to optimize the set of imputation parameters, a simulation was performed on chromosome 1 in 10 replicate (Figure S19). The simulation result on chromosome 1 in 368 lines indicated that the the missing rate was reduced from 91.6% (1–7,818/88,581) to 12.9%, with an accuracy rate 96.6% (Table S3). The optimal parameter combination (IBD: SNPs number≤150 in 5 Mb window size; KNN: w = 20, k = 6, p = −7, r = 1) was then used to impute the missing SNPs for the remaining 145 inbred lines, resulting in an 85.5% filling rate. Therefore, our approach combining SNP-chip data and RNA-seq SNP data with an effective projection procedure permits the quick construction of a high-density physical map and integration of SNPs from RNA-seq data set onto the whole population. This

SNP projection and imputation

The whole panel, 513 maize inbred lines, was genotyped using the MaizeSNP50 BeadChip containing 56,110 SNPs (Illumina). RNA sequencing was performed on immature seeds for 368 out of the 513 maize inbreds using 90-bp paired-end Illumina sequencing, resulting in 2,445.9 Gb of raw sequencing data. 536,809 high quality SNPs obtained by combining the two genotyping platforms (RNA-seq and SNP array) [5,16] were used in the study. For the additional 145 maize lines, the genotype calls of unique loci from the integrated SNP data were projected based on regions of IBDFD to physical maps constructed using 56110 SNPs, and then high-density markers with more than 0.5 million SNPs were obtained for all the lines. Out of 56,110 SNPs from MaizeSNP50 data set, 49728 SNPs overlapped with the integrated SNPs data based on their physical positions (B73 RefGen_v2). The 49,728 common SNPs were regarded as core or frame markers for projection based on IBDFD regions. In order to evaluate the performance of IBDFD [17] based projection, training and validation datasets were established for chromosome 1, which had 7618 core markers from Illumina Maize SNP50 and 88581 SNPs from the integrated data set. The genotypes for one maize line with RNA-seq data in IBDFD regions were assigned to the matched target line without RNA-seq data for each SNP. The projection accuracy was calculated by comparing inferred genotypes of 368 lines with their real genotype obtained from RNA-seq. In addition, KNN algorithm [4] which infers a large number of missing genotypes generated from low-coverage genome sequencing was used to impute the missing genotypes of the unique loci from RNA-seq SNP data based on 49728 frame markers. Single method analysis, either IBDFD based projection or KNN algorithm, cannot achieve both optimal accuracy and coverage (see Materials and Methods). However, the combination of IBDFD based projection and KNN seemed effective to infer a large number of missing genotypes. In order to optimize the set of imputation parameters, a simulation was performed on chromosome 1 in 368 lines (Figure S19). The simulation result on chromosome 1 in 368 lines indicated that the missing rate was reduced from 91.6% (1–7,818/88,581) to 12.9%, with an accuracy rate 96.6% (Table S3). The optimal parameter combination (IBD: SNPs number≤150 in 5 Mb window size; KNN: w = 20, k = 6, p = −7, r = 1) was then used to impute the missing SNPs for the remaining 145 inbred lines, resulting in an 85.5% filling rate. Therefore, our approach combining SNP-chip data and RNA-seq SNP data with an effective projection procedure permits the quick construction of a high-density physical map and integration of SNPs from RNA-seq data set onto the whole population. This

Author Summary

Genotype imputation has been used widely in the analysis of genome-wide association studies (GWAS) to boost power and fine-map associations. We developed a two-step data imputation method to meet the challenge of large proportion missing genotypes. GWAS have uncovered an extensive genetic architecture of complex quantitative traits using high-density SNP markers in maize in the past few years. Here, GWAS were carried out for 17 agronomic traits with a panel of 513 inbred lines applying both mixed linear model and a new method, the Anderson-Darling (A-D) test. We intend to show that the A-D test is a complement to current GWAS methods, especially for complex quantitative traits controlled by moderate effect loci or rare variations and with abnormal phenotype distribution. In addition, the traits associated QTL identified here provide a rich resource for maize genetics and breeding.
Statistical power of the imputation-based association test

The 368 maize inbreds with 556,809 SNPs genotyped by RNA-seq and Maize SNP50 array were defined as Data set 1. In addition, Data set 1 and 145 maize inbred lines with joint IBD-based projection and KNN imputed genotypes were defined as Data set 2 together. The 513 maize inbreds with Maize SNP50 array genotyped were defined as Data set 3. To evaluate the reliability of imputed genotypes for 145 inbred lines, GWAS was performed using MLM, with both Data sets 1, 2 and 3 focusing on kernel oil concentration, which has been thoroughly analyzed in our previous study [5].

For GWAS performed using MLM with both Data sets 1 to 3, a total of 26, 32 and 8 significant loci were identified in Data sets 1 to 3, respectively, at the Bonferroni-corrected threshold ($-\log P > 5.74, \alpha = 1$) (Figure 1). Almost all strong signals identified in data sets 1 and 3 were also identified in data set 2 (Table S4). More interestingly, we identified six additional significantly associated loci in dataset 2 ($-\log P > 5.74, \alpha = 1$), including the phosphoinositide 3-kinases gene (PI3Ks) and the phosphatidylinositol transfer protein, which is known to be involved in the oil concentration trait [18] (Figure 1, Table 1, Table S4). This suggests that GWAS carried out using the imputed genotypes with a larger population size ($n = 368$) or low density as DNA array SNPs with the same population size ($n = 513$) increased the statistical power compared to the analyses of RNA-seq genotyped SNPs with the smaller population size ($n = 513$) increased the statistical power compared to the analyses of RNA-seq genotyped SNPs with the smaller population size ($n = 368$) or low density as DNA array SNPs with the same population size ($n = 513$).

GWAS for 17 agronomic traits using MLM

GWAS for 17 agronomic traits using MLM was conducted with Data set 2 and the results are summarized in Figures S1C, D, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17C, D. A total of 19 significant SNPs from 10 loci were identified for five traits (ear leaf width, ear length, kernel width, plant height and tassel main axis length) (Table 2). No significant SNP was found to be associated with the other 12 tested traits at the Bonferroni-corrected threshold ($-\log P > 5.74, \alpha = 1$). If we set the Bonferroni-corrected threshold as $-\log P > 7.05 (\alpha = 0.05)$, no SNPs were significant for all the 17 traits. It may be too strict to use 0.05/n as the cutoff since not all the markers are independent. One thousand permutation tests were conducted for three typical traits with different level of population structure (kernel width, ear height and days to heading) (Table S2). The results showed the cutoff value at $\alpha = 0.05$ is quite similar (Table S3) with 0.05/n.

Anderson-Darling (A-D) test, an alternative method for GWAS

The A-D test [19] is a nonparametric statistical method and a modification of the KS test [12,20] that gives more weight to the tails of the distribution than the KS test. Since the identified loci were much less numerous than expected using the MLM method, the data set was reanalyzed using the A-D test. The same three traits (kernel width, ear height and days to heading) were used to perform 1000 time permutation tests to determine the cutoff values. The results showed the cut off value at $\alpha = 0.05$ varied around the Bonferroni-corrected threshold as 0.05/n (Table S5). To simplify the procedures, we used the uniform cutoff $-\log P > 7.05, \alpha = 0.05$ for further analysis. Flowering time is an important and well-studied trait, and many QTL or candidate genes have been identified [21,22]. Recently, several studies have confirmed that ZmCCT is the gene underlying the major QTL affecting flowering time on chromosome 10 [22,23]. Taking flowering time as an example, it provides a good opportunity to test whether A-D test is a feasible GWAS method for agronomic traits or not. Using the A-D test, we identified 30 loci associated with days to heading in Yunnan 2010. Around 20% of 30 loci were located within a QTL support interval reported in NAM population [22]. If the significant loci are randomly distributed in the genome, the probability by chance is equal to the ratio between the whole-length of QTL interval and the whole genome length (12%), which represents an almost twofold enrichment compared with the 12% expected by chance. A strong association ($-\log_{10}(P) = 7.59$) was identified in 1.7 Kb upstream of ZmCCT (Figure 2A). Four other loci seem to be strong candidates including: one homologous gene (CIB1) [24] shown to be involved in the regulation of flowering time in Arabidopsis, two homologues containing CCT domain that was demonstrated as key photoperiod regulatory gene in plants
Table 1. Additional SNPs and candidate genes significantly associated with oil concentration found using imputed genotype data.

| Candidate gene | Chr | Position | SNP       | *Alleles | MAF | eQTL | P(n = 368) | P(n = 513) | Annotation                                      |
|----------------|-----|----------|-----------|----------|-----|------|------------|------------|------------------------------------------------|
| GRMZM2G132898  | 1   | 105672248| M1c105672248 | G/T      | 0.07| N.S. | 9.21E-05  | 1.58E-06  | choline-phosphate cytidylyltransferase, CCT   |
| GRMZM2G355572  | 3   | 7938035  | M3c7938035 | A/C      | 0.09| N.S. | 5.39E-06  | 1.25E-06  | Unknown                                         |
| GRMZM2G142315  | 3   | 156963535| M3c156963535 | C/G      | 0.05| 6.70E-12| 5.42E-06  | 8.46E-08  | phosphatidylinositol transfer protein, CSR1   |
| GRMZM2G122277  | 4   | 31245730 | PZE-104026172 | A/C     | 0.11| N.S. | 1.91E-04  | 7.87E-07  | Unknown                                         |
| GRMZM2G138245  | 5   | 22806147 | M5c22806147 | A/G      | 0.06| 8.59E-13| 4.77E-05  | 2.14E-07  | Phosphoinositide 3-kinase, PI3Ks               |
| GRMZM2G467356  | 7   | 139139746| M7c139139746 | C/G      | 0.06| N.S. | 2.93E-05  | 4.75E-08  | Unknown                                         |

*Alleles with underlines indicate rare alleles.
N.S: non-significant.

doi:10.1371/journal.pgen.1004573.t001

Table 2. Comparison of significant SNPs identified for 17 traits from MLM using imputed data (dataset 2).

| Traits | Chr. | Position | SNPs  | Candidate genes | Annotation                                      | −Log10 (P) | PVE (%) |
|--------|------|----------|-------|-----------------|------------------------------------------------|------------|--------|
| EL     | 1    | 50646115 | A/G   | GRMZM2G329040   | Haloacid dehalogenase-like hydrolase            | 6.91       | 5.32   |
|       | 1    | 50668188 | A/C   | GRMZM2G003565   | Thioredoxin-like fold                           | 6.82       | 5.41   |
|       | 1    | 50712263 | T/C   | AC2085714_FG001/GRMZM5G851485 | Six-bladed beta-propeller, Tox1-like/Helix-loop-helix domain | 6.71       | 3.5    |
| TMAL   | 1    | 229858329| T/C   | GRMZM2G079428   | Unknown                                         | 5.76       | 5.52   |
|        | 1    | 285931176| A/G   | GRMZM2G34533    | Protein kinase, catalytic domain                | 6.18       | 5.17   |
| PH     | 3    | 162709488| A/G   | GRMZM2G01050    | Unknown                                         | 6.56       | 9.61   |
| KW     | 7    | 14846475 | A/C   | GRMZM2G13044    | Unknown                                         | 5.83       | 0.03   |
| EL     | 8    | 101507990| A/G   | GRMZM2G164090   | Gibberellin regulated protein                    | 6.05       | 4.23   |
| ELW    | 10   | 126586283| A/G   | GRMZM2G167280   | Protein kinase, catalytic domain                | 5.81       | 5.73   |

*The Bonferroni-corrected threshold is −Log10 (P) > 5.74 (the Bonferroni-corrected thresholds for the P values were 1.796 × 10⁻⁸ and corresponding −Log10 (P) values of 5.74 for 556809 SNPs, at α = 1).

Chromosome; PVE, explained phenotypic variation; EL, ear length; DTH, days to heading; TMAL, tassel main axis length; PH, plant height; ED, ear diameter; KW, kernel width; DTA, days to anthesis; ELW, ear leaf width.
doi:10.1371/journal.pgen.1004573.t002
and one locus previously shown to affect flowering time in maize (Id1) [26] (Figure 2A). Using the MLM method, we were only able to identify the marginally significant association for ZmCCT ($-\log_{10}(P) = 5.64$) and there were no strong signals in other genome regions (Figure 2B). Therefore, A-D test could be a more appropriate GWAS method for agronomic traits and we performed GWAS using the A-D test in each subpopulation of Data set 2 without controlling of population structure for all tested traits. The total number of unique SNPs significantly associated with the 17 traits was 678, of which 310 represented unique loci (Figure S1E, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17E). The numbers of significant SNPs associated with different traits ranged from 1 (Cob Weight) to 35 (Tassel branch number and Days to silking). For plant height, a total of 107 loci were identified at the Bonferroni-corrected threshold ($-\log P > 7.03, \alpha = 0.05$) (Figure S10). About 10% of the loci were detected to affect two or more different traits that were consistent with the observed correlations among the measured traits (Figure S10). There were 101, 71 and 171 loci detected in three subpopulations: SS (subpop-1), NSS (subpop-2) and TST (subpop-3), respectively. A reasonable number of spurious associations should be existed in the detected loci since the population structure is not properly addressed in each subpopulation. Genomic control [27–30] is a popular method to control the population stratification and cryptic relatedness that was applied to adjust A-D test statistic in each subpopulation in present study. In total, 19 loci were significantly associated with 13 traits at the Bonferroni-corrected threshold ($-\log P > 5.74, \alpha = 1$) (Table S7).

To further examine the nature of statistically significant associations, we examined the phenotype distributions of individuals carrying each allele. Interestingly, some associations that differed in the width of the phenotype distribution but which had nearly identical trait means were found to be highly significant by the A-D test but not significant by MLM. Figure 3 illustrated significant loci for ear height with nearly identical trait means (Figure 3A) and significant loci for ear leaf width with an obvious shift of the means (Figure 3B). In total, 14.6% of significant loci identified by A-D test do not have an obvious shift of the mean between the two alleles ($t$-test, $p > 0.05$) (Table S7). In this case the differences between distributions are real, but the corresponding genetic markers would not be useful in breeding if the objective is to change the phenotypic means.

Comparison of different association mapping methods based on simulated data

Causal allele frequencies and trait distributions are the main factors that affect association mapping efficiency [1,31]. GWAS data were simulated by adding phenotypic effects to real genotypic data considering the population structure and epistasis from MaizeSNP50 BeadChip [13] under three scenarios: a normal distribution model, an abnormal distribution model caused by uncertain effectors like phenotyping errors and an abnormal distribution model caused by a larger effect QTN with rare alleles (Methods). We compared our noticed A-D test with three other mapping methods: Kruskal-Wallis (K-W) test and linear model (LM) which does not correct for population structure; MLM which corrects for population structure and kinship. Statistic power of the four methods were compared under the same level Type I error. For each method, QTNs were considered to be detected if their $P$ value were below the threshold determined by 1,000 times permutation.

The results for the three simulation schemes are shown in Figure 4 and can be summarized as follows: First, MLM has more (Figure 4A, C, I) or similar (Figure 4G) power among the four methods for major QTNs in schemes 1 and 3. Second, regardless of the allele frequency, nonparametric methods usually have greater power than LM and MLM for moderate QTNs (Figure 4D–F, J–I). Third, A-D test is more powerful than K-W
test in terms of QTNs with rare alleles (Figure 4J–I). Fourth, nonparametric methods are much more powerful than parametric methods in scheme 2 (Figure 4B, E, H, K) that the phenotype has an abnormal distribution model caused by uncertain effectors.

**Co-localization of QTL and candidate genes for agronomic traits**

We compared our mapping results for 17 agronomic traits with QTL identified using different linkage segregation populations and with previously reported known genes. Loci identified by the A-D test which overlap with previously identified genes and loci mapped in biparental populations are summarized in Table S7. Considering the large confidence interval of previously reported QTL, 3 independent RIL populations genotyped with high-density SNP markers were used to conduct QTL analysis for 3 traits (kernel width, ear length and kernel number per row) of the 17 traits tested. For the compared traits, 9 loci (20% of the detected loci) identified by the A-D test were within the QTL confidence interval. One example is a major kernel width QTL which was mapped in chromosome 7 with BK/Yu8701 RILs and explains 18.7% of phenotypic variation (Figure 5B). Within the QTL interval, significant SNPs-kernel width association was detected (Figure 5A). Six candidate genes: GRMZM2G354539, GRMZM2G052893, GRMZM2G052817, GRMZM2G354525, GRMZM2G052610, and GRMZM2G052509 are found in the associated interval. An expression quantitative trait locus (eQTL) was detected for one (GRMZM2G052509, \(\log_{10}(P) = 10.16\)) of the six annotated genes, which can therefore be regarded as a candidate gene for further study. The second example of overlap included the SNP chr2.s.1972207(C/G) with \(\log_{10}(P) = 9.06\) and SNP chr2.s.1972176(C/G) with \(\log_{10}(P) = 8.49\) which were both significantly associated with ear length, and a QTL affiliated with ear length identified in B73/By804 RIL near the associated peak (Figure 5D–F). SNP chr2.s.1972207 (C/G) and SNP chr2.s.1972176 (C/G) were the only two of the 36 SNPs within the gene GRMZM2G061877, which encodes a DHHC zinc finger domain containing protein, and both of them are in the CDS region. SNP chr2.s.1972176 (C/G) makes no difference to the translated protein sequence, while SNP chr2.s.1972207 (C/G) results in a change of Isoleucine to Methionine. Several zinc finger proteins that play important roles in maize inflorescence development, for instance transcription factors RA1, RA2 and RA3 in ramosa pathways [32], have been identified. In rice, a zinc finger transcription factor DST directly regulates OsCKX2 expression in the reproductive meristem leading to OsCKX2 regulated CK accumulation in the shoot apical meristem (SAM) and, therefore, controls the number of the reproductive organs; the \(dst\) mutant leads to lower plant height and longer rice panicle length [33]. These zinc finger genes are functioning as transcription factors. Since the DHHC protein domain product of GRMZM2G068177, which was strongly suggested as a candidate gene for the regulation of ear length, acts as an enzyme, this may suggest a novel function of zinc finger proteins in monocot reproductive organ development. However, further work is needed to test this hypothesis. The third example, one QTL located on chromosome 1 using K22/Dan340 RIL population which explains 11.8% of phenotypic variation for kernel number per row, also overlapped with significant association signals (Figure 5G–I). Four candidate genes: GRMZM2G088524, GRMZM2G022822, GRMZM2G108180 and GRMZM2G052666, located in a 200 kb window around the significant signals, were predicted.

**Discussion**

The genome-wide imputation of genotypes has attracted much attention given its broad applicability in the GWAS era. There are a number of methods for imputing missing genotypes, but many factors influence the accuracy of imputed genotypes [34,35]. In this study, we proposed a two-step method combining IBD based projection and KNN algorithm to infer missing genotypes, resulting in 96.6% accuracy and 85.5% genome coverage in the tested samples. Considering that the missing genotypes consist of over 91.6% of our raw data set, this level of accuracy is acceptable.
Figure 4. Power comparisons in three simulation schemes for four different mapping methods: A-D, KW, MLM and LM. The “Power” was defined as the detection frequency in 500 repeats for a certain QTN. For the purpose of computing power, a causal SNP was considered to be detected only when the causal SNP was significant at a threshold from 1,000 times permutations. The power and type I error of major QTNs (A–C).
and moderate QTNs (D–F) with common allele frequency. The power and type I error of major QTNs (G–I) and moderate QTNs (J–L) with rare allele frequency. A-D test: Anderson–Darling test; LM: linear model; K-W test: Kruskal-Wallis test; MLM: mixed linear model. Scheme 1, phenotypes with normal distribution; Scheme 2, phenotypes with abnormal distribution caused by uncertain effectors; Scheme 3, phenotypes with abnormal distribution caused by a larger effect QTN with rare allele frequency. doi:10.1371/journal.pgen.1004573.g004

Compared with other methods [4,17,34,35], the two-step method has its advantages. Imputation based only on IBD regions ensures high accuracy but a relatively low coverage rate. KNN algorithm has been proven to be a good strategy for sequencing data [4], however, it alone does not represent the true similarity of the inbred lines due to low density of frame markers and rapid LD decay in maize in our study. Therefore, we first used IBD based imputation to increase marker density, and then the KNN algorithm was used to infer the missing genotype, leading to high coverage rate and imputation accuracy. Imputation error is often caused by ignoring recombination and mutation within IBD regions. In addition, if an inbred line with low density markers share a region with two or more inbred lines with high density markers and the missing genotypes are inferred on the basis of only one of these lines, there is a high risk of error since the accuracy of the projection depends on the identity between the projected and chosen lines. Reanalysis of GWAS for kernel oil concentration revealed consistent results and a higher detection power. Six more associated loci were identified, most likely due to the increase in sample size. The implication is that mapping resolutions are enhanced by extracting moderately more information from the genome and expanding sample size.

The detection of loci controlling complex traits using GWAS has flourished and numerous statistical approaches for GWAS analysis in plants have recently been described [14,36–40]. Linear statistical models like ANOVA, general linear model (GLM), and MLM establish significance cutoff by relying on the assumption that target traits have normal distribution. However, sometimes phenotype distribution in the moderate plant population is not normal in the tails that may be due to the population size, field experiment such as phenotyping errors, or genetic effects [31]. Based on our simulated data, nonparametric methods including A-D and KW tests usually have greater power than LM and MLM for abnormal phenotypes, rare alleles and moderate QTNs. It also implies that A-D and KW tests should perform well to detect the shifts of distribution as well as changes in the shape of distributions [31]. A-D test possesses advantages than K-W test in the detection of QTNs with rare alleles. However, MLM performs better than A-D and KW test for the major QTNs especially those with common alleles. However, we need to keep in mind that population structure of the studied samples is the key confounder for GWAS. In the measured 17 agronomic traits of present study, we observed the phenotypic variation explained by population structure ranged between 0.9% and 32.3% (Table S2). In the A-D test, we didn’t account the confounding by population structure in the subpopulation that may lead to false-positive findings. Genomic control is a good alternative for controlling the statistic inflations [27–30], different inflation factors were observed in different traits and different subpopulations in present study (Table S7). We detected 19 loci significantly associated with 13 traits at the Bonferroni-corrected threshold (-log P>3.74, $\alpha = 1$) (Table S7) using genomic-control ($\hat{\theta}_{\text{regress}}$) to adjust our real phenotype test statistic from A-D test. However, we also need to be careful that the adjusted -Log P might be over corrected, since A-D test has already controlled part of the population structure and genomic control method is affected significantly by the true association signals, even for the agronomic traits may involve a larger number of loci with small effects [21,36,37,41]. And the influence of epistatic genetic effect to the genomic control is still not explored [27–30]. Another thing need to be noticed is testing within subpopulations (A-D test) and across the whole panel with controlling the population structure (MLM) are different. Testing within subpopulations changes allele frequencies of background alleles and therefore possibly changes the epistatic interactions that are mapped in an additive manner within subpopulations but were not mapped across populations.

In general, A-D test could be a good complement to current popular GWAS methods. As each method owning its own advantages, the preliminary understanding of the traits studied is needed for choosing GWAS methods or trying different GWAS methods would be helpful especially for those studies only few or none significant signals were identified by using only one method. In this study, we performed GWAS using both MLM and A-D test for 17 agronomic traits. In total, 18 overlapped regions were detected by the two approaches (Table S7). The A-D test also showed high concordance with previous studies in identifying a higher number of QTL related to agronomic traits.

Our noticed nonparametric statistical approach is robust with respect to non-normality, similarly to the KS test [12]. The KS test tends to be more sensitive around the median value and less sensitive at the extreme ends of the distribution. Thus, the KS test is not always appropriate for calculating the significance of data sets which differs at the tails of the probability distribution, while the median remains unchanged [31]. The A-D test improves upon the WS test because it has more sensitivity towards the tails of the pooled sample. More importantly, the performance of the A-D test for small samples is quite good, as demonstrated by numerous Monte Carlo simulations [19]. This means that, for complex traits, the A-D test can make a good use of SNPs that have minor allele frequency and keep detection ability to the relatively small effect loci. At the same time, it is important to recognize that there are always limitations to what can be achieved using statistics. It seems that A-D test does not work well for all traits. Interestingly, we identified 14.6% associations by A-D test that differed in the width of the phenotype distribution but which had nearly identical trait means (Figure 3A). In these cases the differences between distributions are real, but the corresponding genetic markers would not be useful in breeding if the objective is to change the phenotypic means. Instead, the associations appear to represent allelic differences in the apparent trait stability. Therefore, to confirm candidate loci, it is necessary to check both frequency distribution and normality of the distribution curves (Figure 3). Several studies in humans have confirmed that using multiple methods for statistical inference critically enables the interpretation of results and engenders stronger candidates for experimental follow-up [42].

We identified some genes affecting important agronomic traits in maize that are very good candidates for future detailed analysis, for allele mining to identify functional variation, and for marker development. As whole genome sequences become available for many crop species including maize, as well as for multiple genotypes of the same species through resequencing, along with cost-effective high-throughput genotyping systems and the next generation of sequencing technologies, GWAS becomes practical and its use in plant breeding will allow the manipulation of many traits at the whole-genome level. Association mapping using a set
of global diverse breeding germplasm and high-throughput SNP markers, as shown in this study, provides high-resolution dissection of the genetic architecture of complex traits. This knowledge in turn will be useful not only for designing marker-assisted selection strategies but also for optimizing conventional breeding systems.

**Materials and Methods**

**Plant material and phenotyping**

A total of 513 maize lines with tropical, subtropical and temperate backgrounds representing the global maize diversity were employed for genome-wide association mapping in this study. All maize inbred lines have been well described in previous studies [13,15] and the 513 maize lines were classified into four subgroups based on population structure Q matrix: Stiff stalk (SS) with 112 lines, Non-stiff stalk (NSS) with 116 lines, Tropical-subtropical (TST) with 258 lines, and an admixed group with 27 lines (detailed information also can be downloaded at (www.maizego.org/resource). A Randomized complete block design with one to two replications was used for field trials in five environments, including Ya’an (30°N, 103°E), Sanya (18°N, 109°E), Yunnan (25°N, 102°E) in 2009, Guangxi (23°N, 110°E) and Yunnan (25°N, 102°E) in 2010. A row length of 3 m was used for each line including 11 plants plot–1 with 25 cm plant to plant and 60 cm row to row distance. Five randomly selected plants were used for phenotypic data acquisition in each line and the mean data in each replication was used for phenotypic analysis. A total of 17 economically important traits were phenotyped (Table S1). These traits were divided into three categories: morphological attributes (plant height, ear height, ear leaf width and length, tassel main axis length, tassel branch number, and leaf number above ear), yield related traits (ear length and diameter, cob diameter, kernel number per row, 100-grain weight, cob weight and kernel width), and maturity traits (days to heading, anthesis, and silking).

**Imputation yield and accuracy**

Imputation methods have not been developed to deal specifically with low density of SNP marker data. Of the available imputation models, identity by descent (IBD) based projection [17] and the k-nearest neighbor algorithm (KNN) [4] seemed to effectively infer a large number of missing genotypes. To assess the performance of IBD based projection, preliminary tests for chromosome 1 in 368 maize lines were conducted. We removed genotype data without frame SNPs and then compared the observed genotypes with those generated by projection. The number of IBD regions with consecutive SNPs for 368 lines varied from 1 to 285 on chromosome 1, and projection accuracy, defined as the percentage of correctly projected genotypes ranged from
QTL mapping

The linkage analyses of ear length, kernel number per row, and kernel width were performed in three recombination inbred line (RIL) populations, BY804/B73 (197 individuals), K22/Dan340 (197 individuals), and BK/Yu8701 (165 individuals). All the RIL lines and their parents were genotyped using Maize SNP50 assay (Illumina) containing 56,110 SNPs [14]. The phenotype of BK/Yu8701 in Henan 2011 and BLUP value from 5 environments of BY804/B73 and K22/Dan340 were used. QTL mapping using the composite interval mapping method [43] was performed in the package QTL cartographer version 2.5 [44].

Statistical analysis and association mapping

ANOVA, correlation, and repeatability analyses for 17 agronomic traits were conducted using SAS software (Release 9.1.3; SAS Institute, Cary, NC). Heritability analysis and association analysis for the 17 agronomic traits in Data set 2 were conducted by MLM using TASSEL [45] software package. The observed p values from marker-trait associations were used to display Q-Q plots and Manhattan plots, using R. Permutation tests were used to determine the cutoff for GWAS. Considering the computation time, we only choose three typical traits with different population structure effects (kernel width, ear height and day of flowering time) as examples. The results showed that the cutoff values are similar with the Bonferroni correction. To simplify the procedures, we use the uniform Bonferroni-corrected thresholds at $\alpha = 1$ and $\alpha = 0.05$ as the cutoffs. When performing n tests, if the significance level for the entire series of tests is $\alpha$, then each of the tests should have a probability of $P = \alpha/n$. When the numbers of markers was 556,089 SNPs, at $\alpha = 1$ and $\alpha = 0.05$, the Bonferroni-corrected thresholds for the $p$ values were $1.796 \times 10^{-6}$ and $8.95 \times 10^{-6}$, with corresponding $-\log p$ values of 5.74 and 7.05, respectively. Regression estimator ($\hat{\beta}_{\text{regress}}$) of Genomic Control inflation factor was used [20]. Percentage of PVE by associated SNPs was calculated by ANOVA. Informative SNPs and candidate genes at the identified loci for the corresponding traits were from public maize genome data set B73 RefGen_v2.
Simulations of the phenotypes were repeated 500 times in all schemes. All simulated phenotypes had been analyzed with the four methods presented in the main text. 1,000 permutations had been done separately for the four methods to obtain the threshold at different type I error risk.

Anderson-Darling test

The Anderson-Darling two-sample procedure assumes that the two samples have a continuous distribution function and we are interested in testing the null hypothesis that the two phenotype samples divided by two alleles of one SNP have the same distribution, without specifying the nature of population:

\[ H_0 : F_1 = F_2 \]

The test procedure is as follows:

1. Calculate \( A_{2KN}^2 \):

The computational formula for \( A_{2KN}^2 \) not adjusted for ties is,

\[
A_{2KN}^2 = \frac{1}{N} \sum_{i=1}^{k} \left( \frac{1}{n_i} \sum_{j=1}^{L-1} \left( \frac{(nF_j - j)H_j^2}{j(N-j)} \right) \right)
\]

and the corresponding adjusted for ties is,

\[
A_{2KN}^2 = \frac{N-1}{N^2} \sum_{i=1}^{k} \left( \frac{1}{n_i} \sum_{j=1}^{L} \left( \frac{(NF_j - nH_j)^2}{H_j(N-H_j) - Nh_j/4} \right) \right)
\]

where:

- \( F_1, F_2 \) indicates the two phenotype distribution function
- \( k = 2; i = 1, 2 \)
- \( n_i = \) data number in the \( r \)th sample; \( j = 1, 2, \ldots, n_i \)
- \( N = \) total number of two samples’ individuals; \( N = n_1 + n_2 \)
- \( x_j = \) data in the \( s \) sample and \( j \) observation within that sample
- \( L = \) the number of unique data, where it will be less than \( n \) with tied data
- \( z[j] = \) distinct values of all combined data ordered in ascending way denoted \( z[1], z[2], \ldots, z[L] \)
- \( h_i = \) number of values in the pooled sample equal to \( z[j] \)
- \( H_j = \) number of values in the combined samples less than \( z[j] \) plus one half of the number of values in the combined samples equal to \( z[j] \)
- \( F_j = \) number of values in the \( r \)th sample which are small than \( z[j] \) plus one half the number of values in this sample which are equal to \( z[j] \)

2. Calculate \( \sigma_N^2 \):

Under \( H_0 \), the variance of \( A_{2KN}^2 \) is,

\[
\sigma_N^2 = \text{var}(A_{2KN}^2) = \frac{aN^3 + bN^2 + cN + d}{(N-1)(N-2)(N-3)}
\]

with:

\[
a = (4g - 6)(k - 1) + (10 - 6g)S
\]

\[
b = (2g - 4)k^2 + 8Tk + (2g - 14h - 4)S - 8T + 4g - 6
\]

\[
c = (6T + 2g - 2)k^2 + (4T - 4g + 6)k + (2h - 6)S + 4T
\]

\[
d = (2T + 6)k^2 - 4Tk
\]

where:

\[
S = \sum_{j=1}^{k} \frac{1}{n_j},
\]

\[
T = \sum_{j=1}^{N-1} \frac{1}{j}^2, R = \sum_{j=1}^{N-2} \sum_{j=1}^{N-1} \frac{1}{(N-j)}
\]

3. Calculate \( T_{KN} \):

\[
T_{KN} = \frac{A_{2KN}^2 - (k-1)}{\sigma_N^2}
\]

4. Refer \( T_{KN} \) to the upper \( \alpha \) percentiles \( T_m(\alpha) \) of the \( T_m \) distribution table below, reject \( H_0 \) at significance level \( \alpha \) if \( T_{KN} \) exceeds the given point \( T_m(\alpha) \). If \( T_{KN} \) is outside the range of the table, plotting the log-odds of \( \alpha \) versus \( T_1(\alpha) \), a strong linear pattern indicates that simple linear extrapolation should give good approximate \( p \) values.

\[
\alpha = 0.25, 0.10, 0.05, 0.025, 0.01
\]

\[
t_1(\alpha) = 0.326, 1.225, 1.960, 2.719, 3.752
\]

\[
m = k - 1
\]

**URL**. One R package (ADGWAS) for GWAS by Anderson-Darling test can be downloaded here: http://www.maizego.org/Resources.html

**Supporting Information**

**Figure S1** Genome-wide association analysis of plant height. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. Know genes controlling the traits were labeled. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for plant height. The three subpopulations: SS (subpop-1), NSS (subpop-2) and TST (subpop-3). (TIF)

**Figure S2** Genome-wide association analysis of ear height. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for ear height. (TIF)

**Figure S3** Genome-wide association analysis of ear leaf width. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for ear leaf width. (TIF)

**Figure S4** Genome-wide association analysis of ear leaf length. (A, B) Phenotype histogram and distribution of subpopulations in
513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for ear length.

(TIF)

Figure S5  Genome-wide association analysis of tassel main axis length. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. Know genes controlling the traits were labeled. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for tassel main axis length.

(TIF)

Figure S6  Genome-wide association analysis of tassel branch number. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. Know genes controlling the traits were labeled. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for tassel branch number.

(TIF)

Figure S7  Genome-wide association analysis of leaf number above ear. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for leaf number above ear.

(TIF)

Figure S8  Genome-wide association analysis of ear length. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. Know genes controlling the traits were labeled. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for ear length.

(TIF)

Figure S9  Genome-wide association analysis of ear diameter. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. Know genes controlling the traits were labeled. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for ear diameter.

(TIF)

Figure S10  Genome-wide association analysis of cob diameter. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. Know genes controlling the traits were labeled. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for cob diameter.

(TIF)

Figure S11  Genome-wide association analysis of kernel number per row. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. Know genes controlling the traits were labeled. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for kernel number per row.

(TIF)

Figure S12  Genome-wide association analysis of 100-grain weight. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. Know genes controlling the traits were labeled. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for 100-grain weight.

(TIF)

Figure S13  Genome-wide association analysis of cob weight. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. Know genes controlling the traits were labeled. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for cob weight.

(TIF)

Figure S14  Genome-wide association analysis of kernel width. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for kernel width.

(TIF)

Figure S15  Genome-wide association analysis of days to anthesis. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. Know genes controlling the traits were labeled. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for days to anthesis.

(TIF)

Figure S16  Genome-wide association analysis of days to silking. (A, B) Phenotype histogram and distribution of subpopulations in 513 maize lines. (C) Manhattan plots of mixed linear model conducted in imputation data, respectively. (D) Quantile-Quantile plots of p-values of mixed linear model conducted in imputation data. (E) Summary of GWAS results from Anderson-Darling test performed on each subpopulation independently for days to silking.

(TIF)

Figure S17  Genome-wide association analysis of days to heading. (A, B) Phenotype histogram and distribution of
Population structure.

The percentage of phenotypic variance explained by the subpopulations of maize lines. a ANOVA, analysis of variance, showing the mean and lines were fitted in the model as random effects. ** indicate applied to determine the significance level. Both the environments square and degrees of freedom (in parentheses). The F-test was validation dataset.

(TIF)

Table S18 Pair-wise Pearson’s correlation among 17 traits in 513 maize lines.

(TIF)

Table S19 The mean error ratio and mean coverage ratio pool over SNP number within IBD region (A) and size of IBD region (B), respectively, on chromosome 1 for the 368 maize lines. (C) Imputation accuracy and filling rate for each of 72 combinations of variables of KNN. The combination, indicated by arrow, was chosen for final data imputation.

(TIF)

Table S1 Description of the traits evaluated in the study.

(XLSX)

Table S2 Phenotype variation of 17 agronomic traits in 513 maize lines. * ANOVA, analysis of variance, showing the mean square and degrees of freedom (in parentheses). The F-test was applied to determine the significance level. Both the environments and lines were fitted in the model as random effects. ** indicate significance at level of 0.001; s.d., standard deviation. 3PVE by Q, the percentage of phenotypic variance explained by the subpopulation structure.

(XLSX)

Table S3 IBD base projection and KNN imputation for validation dataset.

(XLSX)

References

1. Yan J, Warburton M, Crouch J (2011) Association mapping for enhancing maize (Zea mays L.) genetic improvement. CROP Sci 51: 433–449.
2. Zhu C, Gore MA, Buckler ES, Yu J (2008) Status and prospects of association Mapping in plants. Plant Genome 1: 5–20.
3. Arelis S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, et al. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627–631.
4. Huang X, Wei X, Sang T, Zhao Q, Feng Q, et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42: 961–967.
5. Li H, Peng Z, Yang X, Wang W, Fu J, et al. (2013) Genome-wide association study dissect the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45: 43–50.
6. Zhang Z, Erez O, Lai CQ, Toddhunter RJ, Tiwari HK, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 353–360.
7. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.
8. Zhao K, Turgay CW, Ezezga GC, Wright MH, Ali ML, et al. (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat Commun 2: 467.
9. Li-Beisson YH, Shorrosh B, Beisson F, Anderson MX, Arondel V, et al. (2010) Acyl-lipid metabolism. The Arabidopsis Book 8: e0133, doi/10.1199/tab.0133.
10. Schoff F, Peters M (1987) K-sample Anderson–Darling tests. J Am Stat Assoc 82: 918–924.
11. Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1992) Numerical recipes in C: the art of scientific computing (second edition). Cambridge University Press. Pp. 626–627.
12. Buckley ES, Holland JB, Bradbury PJ, Archarya CB, Brown PJ, et al. (2009) The genetic architecture of maize flowering time. Science 325: 714–718.
13. Hung HY, Shamoon LM, Tian F, Bradbury PJ, Chen C, et al. (2012) ZmCCT and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proc Natl Acad Sci U S A 109: 1913–1921.
14. Yu L, Xu Y, Li K, Kleinpjot J, Yang H, et al. (2008) Photocycle-driven Cry2 interacts with CIB1 to regulate transcription and floral initiation in Arabidopsis. Science 322: 1535–1539.
15. Cockram J, Thi t, Steuerwandel B, Stein N, Taudien S, et al. (2012) Genome dynamics explain the evolution of flowering time CCT domain gene families in the Poaceae. PLoS One 7: e55907.
16. Colasanti J, Tremblay R, Wong AY, Coneva V, Koszak A, et al. (2006) The maize INDETERMINATE1 flowering time regulator defines a highly conserved zinc finger protein family in higher plants. BMC Genomics 7: 158.
27. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
28. Ting Yan, Bo Hou and Yaning Yang (2009) Correcting for cryptic relatedness by a regression-based genomic control method. BMC Genet 10:78.
29. Zheng G, Ferdinun B, Li Z, Gastwirth JL (2005) Genomic control for association studies under various genetic models. Biometrics 61:186–192.
30. Tsepitou V, Ried JS, Strauch K, Grallert H, van Duijn CM, et al. (2013) Development and application of genomic control methods for genome-wide association studies using non-additive models. PLoS One 8:e61431.
31. Belo A, Lack SD (2010) Association mapping for the exploration of genetic diversity and identification of useful loci for plant breeding. In: Khalid M, Gunther K, editors. The handbook of plant mutation screening: mining of natural and induced alleles. New York: John Wiley & Sons. Pp. 231–246.
32. Gallavotti A, Long JA, Stanfield S, Yang X, Jackson D, et al. (2010) The control of axillary meristem fate in the maize ramosa pathway. Development 137:2849–2856.
33. Li S, Zhao B, Yuan D, Duan M, Qian Q, et al. (2013) Rice zinc finger protein DST enhances grain production through controlling Gna1/OsCKX2 expression. Proc Natl Acad Sci U S A 110: 3167–3172.
34. Daetwyler HD, Wiggans GR, Hayes BJ, Woodlana JS, Goddard ME (2011) Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189: 317–327.
35. Hao K, Chudin E, McElroy J, Schadt E (2009) Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genet 10: 27.
36. Kump KL, Bradbury PJ, Wiser RJ, Buckler ES, et al. (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet 43: 163–168.
37. Poland JA, Bradbury PJ, Buckler ES, Nelson RJ (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. Proc Natl Acad Sci U S A 108: 6893–6898.
38. Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, et al. (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3: e4.
39. Stich B, Mohring J, Piepho H-P, Heckenberger M, Buckler ES, Melchinger AE (2008) Comparison of mixed-model approaches for association mapping. Genetics 178: 1743–1754.
40. Stich B, Melchinger AE (2009) Comparison of mixed-model approaches for association mapping in rice and Arabidopsis. BMC Genomics 10: 94.
41. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, et al. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43: 159–162.
42. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78.
43. Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136: 1457–1467.
44. Wang S, Baten CJ and Zeng ZB (2012). Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC.
45. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, et al. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2635–2635.