Chapter 5

Interpretation of Manhattan Plots and Other Outputs of Genome-Wide Association Studies

Jiabo Wang, Jianming Yu, Alexander E. Lipka, and Zhiwu Zhang

Abstract

With increasing marker density, estimation of recombination rate between a marker and a causal mutation using linkage analysis becomes less important. Instead, linkage disequilibrium (LD) becomes the major indicator for gene mapping through genome-wide association studies (GWAS). In addition to the linkage between the marker and the causal mutation, many other factors may contribute to the LD, including population structure and cryptic relationships among individuals. As statistical methods and software evolve to improve statistical power and computing speed in GWAS, the corresponding outputs must also evolve to facilitate the interpretation of input data, the analytical process, and final association results. In this chapter, our descriptions focus on (1) considerations in creating a Manhattan plot displaying the strength of LD and locations of markers across a genome; (2) criteria for genome-wide significance threshold and the different appearance of Manhattan plots in single-locus and multiple-locus models; (3) exploration of population structure and kinship among individuals; (4) quantile–quantile (QQ) plot; (5) LD decay across the genome and LD between the associated markers and their neighbors; (6) exploration of individual and marker information on Manhattan and QQ plots via interactive visualization using HTML. The ultimate objective of this chapter is to help users to connect input data to GWAS outputs to balance power and false positives, and connect GWAS outputs to the selection of candidate genes using LD extent.

Key words GWAS, Linkage disequilibrium, Population structure, Kinship, False positive rate, Mixed linear model

1 Introduction

The genome-wide association study (GWAS) is an important tool to map the genes underlying complex traits in animals and plants, as well as genetic diseases in humans [1–6]. The most common genetic markers, single-nucleotide polymorphisms (SNPs), are used to capture the linkage disequilibrium (LD) with quantitative trait loci (QTL) [7–10]. The detection ability of GWAS is dependent on many factors, including population size, heritability, marker density, minor allele frequency (MAF), statistical model, and genetic architecture of the trait [11–15]. With the

Davoud Torkamaneh and François Belzile (eds.), *Genome-Wide Association Studies*,
Methods in Molecular Biology, vol. 2481, https://doi.org/10.1007/978-1-0716-2237-7_5,

development of sequencing technology, deep sequencing and large populations have been employed for GWAS detection in more and more species [16–22]. Although the associated markers only explain a small proportion of heritability, the proportion is expected to increase with additional discoveries [23–27].

Multiple software packages (e.g., TASSEL [28], GCTA [29], PLINK [30], and GAPIT [31]) have been developed to conduct GWAS. The models used by these packages differ in statistical power, computing speed, and output. The two major outputs of GWAS are Manhattan and Quantile–Quantile (QQ) plots [32–34]. In this chapter, we first describe the various forms of these plots, including subtle differences in presentation due to the selected models (single-locus vs. multiple-locus models). Secondly, we describe the plots that are helpful to determine genotypes, population structure, and familial relatedness (i.e., kinship). Finally, we describe interactive visualization based on application of HTML.

2 Generating Manhattan Plots

The Manhattan plot is used to summarize and visualize the marker-trait relationships across the whole genome. It gained its name from the similarity of such a plot to the Manhattan skyline: a profile of skyscrapers towering above the lower level "buildings" which vary around a lower height [3]. Generally, the *x*-axis denotes the physical positions of the markers on the chromosomes, and the *y*-axis is the negative $\log_{10}$ *P*-values. These *P*-values correspond to the hypothesis test of H_0: There is no association between the tested marker and the studied trait. The tall "buildings" of Manhattan plots represent SNPs with strong statistical association with the tested phenotype. A line typically spans the Manhattan plot from left to right; that named cutoff line is used to highlight markers over a threshold value that are statistically significantly associated with the trait (Fig. 1). Ideally, these thresholds will be calculated by controlling for multiple testing on a genome-wide scale. The significant points and continuous peak of markers are of interest, because they may signify potential nearby QTLs [35, 36]. If there are no markers over the cutoff line, there may be several reasons, such as limited population size, sequencing depth level, or insufficient statistical method.

2.1 P-values and Their Negative Log Transformation

The null hypothesis of an association test is that there is no linkage disequilibrium between a genetic marker and a trait of interest. GWAS tests the hypothesis for all genetic markers across the genome either by testing the markers one at a time in single-locus model [37], or testing all the markers simultaneously [38–40]. There is another type of test between the two extremes

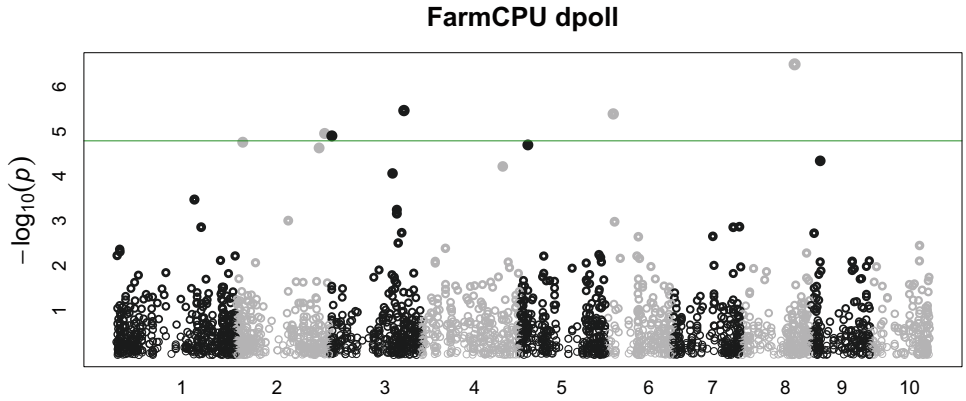

Fig. 1 Manhattan plot of genome-wide association study conducted using FarmCPU. The data contains 281 maize inbreds phenotyped on flowering time (days to pollination) and genotyped with 3093 SNPs

named the multiple loci models [41–43]. Besides testing markers one at a time, multiple loci models also include additional markers as cofactors. The typical statistical tests include *F*-test and *t*-test, which generate probability (*P*-values) for each marker. A high *P*-value suggests a high chance that the null hypothesis is true, otherwise reject the null hypothesis when the *P*-value is below a threshold. To visualize the association across the genome, the *P*-values are plotted as the negative log scale with a base of 10. Therefore, the associated markers appear at the top and the nonassociated markers at the bottom, formatting the well-known Manhattan plot [44].

2.2 Chromosome and Position Visualization

The physical location of markers on a chromosome can be used to indicate the relative position. It is easy to plot a "chromosome-wise" Manhattan plot for each chromosome. However, the physical locations of markers between chromosomes are not continuous. We need to convert the physical locations of markers to continuous numeric relative positions for a whole-genome Manhattan plot. A lot of software or R packages, such as GAPIT, rMVP [45], qqman [46], and Haploview [47], code the relative positions of markers in the next chromosome equal to the each physical location of markers in the next chromosome plus the maximum value of relative position of markers in the previous chromosome. For a polyploid genome, such as allopolyploid wheat, the chromosome names will be coded as consecutive numbers to show the relative position.

2.3 File Size Reduction

Following the development of sequencing technology and the reduction of sequencing cost, genome sequencing depth and density have been improved. More and more large genotype datasets have been used to analyze the association with traits [48–52]. As markers are displayed in log scale based on their *P*-values, most of the markers are on the bottom of the plots on top of each other. Displaying all the markers on the bottom is not only unnecessary,

but also creates problems for storage and display. For example, two million markers will result a PDF file with size over 200 Mb. Therefore, an effective reduction of marker for display is necessary. By default, GAPIT only displays 5000 markers. The algorithm barely removes markers on the top of Manhattan plot. The markers on the bottom are randomly selected to display based on sampling with uniform distribution on the log *P*-values. Consequently, the reduction generates identical visual plot as plotting all markers.

2.4 Circle vs. Cartesian

Comparing linkage analysis based on genetic information, GWAS is based on statistical information. The associated markers across correlated traits provide a partial evidence for their linkage with a causal pleiotropic mutation. Although the causes other than linkage (e.g., population structure) cannot be excluded, the risk of other causes can be dramatically reduced, including phenotypes with outliers and genotypes with rare variants. Researchers are also interested to compare different analyses such as using different statistical models. Therefore, it is beneficial to organize multiple Manhattan plots together and demonstrate the overlaps among them. The plural Manhattan plot has two types: Circle and Cartesian (Fig. 2). The Circle and Cartesian Manhattan plots are available through the software GAPIT [31, 53], rMVP [45], and CMplot [45]. In GAPIT, a marker is indicated by vertical dashed lines if it appears as an associated mark in exactly two single Manhattan plots, or a solid line if it appears in three or more single Manhattan plots.

3 Determining a Significance Threshold to Declare Association

In a statistical test, the type I error is the probability of falsely rejecting the true null hypothesis. In GWAS, the null hypothesis is that the tested marker is not associated with the trait. If the *P*-value for the test statistic at a given marker is less than a threshold (α), we declare this test as significant under this threshold cutoff [54–56]. Conventionally, the threshold equals to 0.05 as significant and 0.01 as very significant to reject the null hypothesis.

3.1 Multiple Test Correction

There are millions of markers in the genome, which means millions of tests. Using a comparison-wise type I error rate of $\alpha = 0.05$, the probability of making at least one type I error across these millions of markers will be substantially larger. This broad and nonstringent threshold cutoff brings huge risk for the detection of false-positive candidate genes [34, 57–59]. The more markers are tested in GWAS, the more erroneous signals are likely to occur. Thus, there is a critical need to adjust for multiple testing. Given below are several commonly used approaches to adjust for multiple testing that are statistically rigorous.

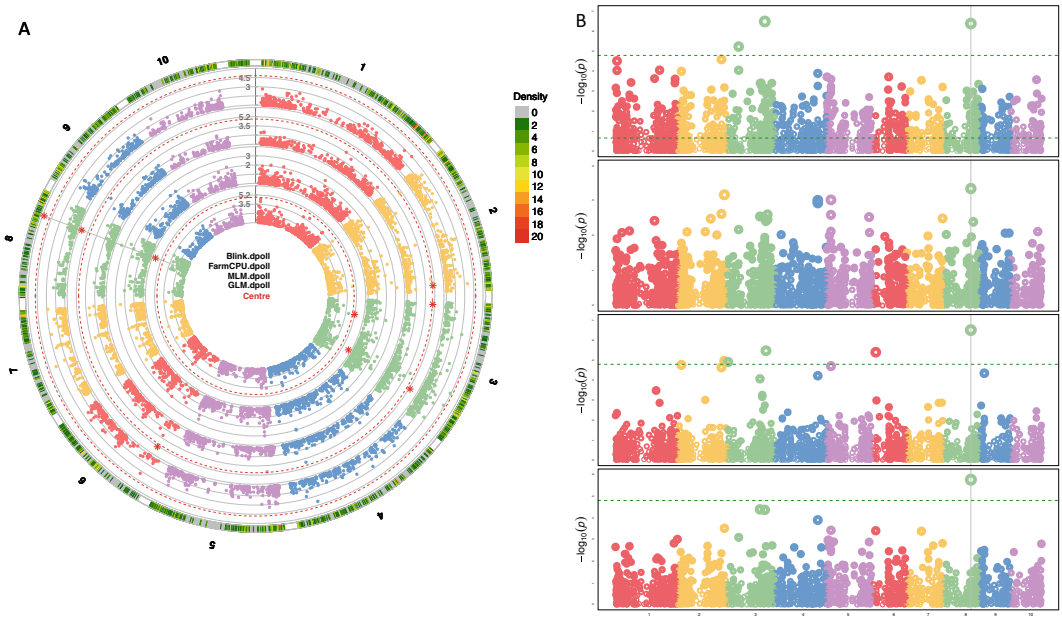

Fig. 2 Manhattan plots of genome-wide association studies using multiple methods. The Manhattan plots are displayed in two formats: Circle (**a**) and Cartesian (**b**). The data consists of 281 maize inbreds phenotyped on flowering time (days to pollination) and genotyped with 3093 SNPs. Four methods were used: BLINK, FarmCPU, MLM, and GLM. Multiple traits can be displayed similarly

3.2 Bonferroni Cutoff

In 1936, Italian mathematician Carlo Bonferroni developed a correction for multiple comparisons using the number of tests and the Type I error for each individual hypothesis [60, 61]. The risk of the error in the single statistical test was repeated multiple times (m). In order to retain a prescribed family-wise error rate (FWER) in an analysis involving more than one comparison, the error rate for each comparison must be more stringent than α. If each of m tests is performed with a type I error rate α/m, the total error rate will not exceed α. The Bonferroni correction could be over-conservative for GWAS, as it assumes independence among multiple comparisons. The markers within a chromosome are in linkage disequilibrium (LD), which contradicts this assumption [6, 62].

3.3 False Discovery Rate (FDR) Cutoff

Bonferroni's multiple test correction uses $\frac{\alpha}{m}$ as the family-wise error rate threshold where α is the type I error threshold of single test and m is number of independent tests. In 1986, Simes proposed an alternative for multiple dependent tests. To maintain the same family-wise error rate α, the number of null hypotheses that can be rejected is k so there are k P-values that are smaller than $\frac{k\alpha}{m}$ [63]. In 1995, Benjamini and Hochberg defined the term as false discovery rate [64]. False discovery rate is the proportion of errors committed by falsely rejecting real null hypotheses. In the

Benjamini and Hochberg procedure, hypotheses are sorted by their *P*-values in ascending order. The hypotheses and corresponding *P*-values are denoted as $H_{(i)}$ and $P_{(i)}$ respectively, $i = 1$ to m. Hypothesis i is rejected if $P_{(i)} \leq \frac{i}{m}\alpha$. The FDR controlling procedures have been proven to have greater power at the cost of increased numbers of Type I errors [14, 64, 65]. In GAPIT, there are two thresholds with family-wise error rate of 0.05 in the Manhattan plot. The solid and dashed green lines stand for the *P*-values corresponding to the Bonferroni and FDR multiple test corrections.

3.4 Permutation Cutoff

The Bonferroni cutoff is overconservative for the genome LD relationship, and the FDR cutoff also may not be restrictive enough for some populations or traits. Referred to as the gold standard GWAS cutoff, the permutation cutoff uses random assortment between genotype and phenotype to derive an empirical cutoff [56, 66, 67]. GWAS is conducted on phenotypes that are randomly shuffled to break the connections with the genotypes. The smallest *P*-value is recorded for each random setting. Multiple replicates (usually more than 100) are required to derive the distribution of the smallest *P*-value for a single replicate. The *P*-values corresponding to the α percentile is the empirical threshold for the family-wise error rate of α. Because a separate permutation procedure is conducted on each trait, there is an individual cutoff established for each individual trait. The disadvantage of this approach is that it requires significant computing time for each trait in a large population.

4 Highlighting Associated Loci

Additional information is usually added to Manhattan plots to assist interpretation. These include, but are not limited to, the color, type and size of dots, the relationship between the top significant marker and neighboring markers, display of known genes or simulated quantitative trait nucleotides (QTNs), and the supplemental LD information. Here we introduce several of these indicators.

4.1 Dot Size and Fill Percentage

Usually, the dot color and type are used to distinguish the markers on the different chromosomes. Some special colors often are used to mark the significant signals. The dots are normally the same size, but for large datasets or high-density markers, the dot sizes are drawn from small to big, based on the markers' *P*-values. That makes the significant markers more conspicuous in the whole Manhattan plot. Also, the opacity of colors can be used to show the marker density. In some simulation GWAS, the positions of real QTNs are known, so the markers are usually represented as circles, and the known QTNs are added as solid points.

4.2 Linkage Disequilibrium with Neighboring Markers

The marker with the strongest association with a trait of interest may not be on a gene. Researchers usually need to inspect the region upstream and downstream of an associated marker to identify candidate genes. Linkage Disequilibrium (LD) between the most associated marker and its neighbors are helpful to determine the candidate genes [16, 26, 68–70]. The chromosome view of the Manhattan plot is used to show not only the details of the association, but also the LD among markers. Heatmap is used to indicate the LD between the most significant marker and its neighbors (Fig. 3).

4.3 Manhattans in New York City (NYC) vs. Kansas

The name of the Manhattan plot is derived from the full skyline view of NYC. The concentrated distribution of the markers' *P*-values looks like tall buildings. However, there is a small town in Kansas also named Manhattan. In the skyline view of this small town, most of the buildings are not tall. The highest human-made object may be the helicopter in the sky. That is similar to some

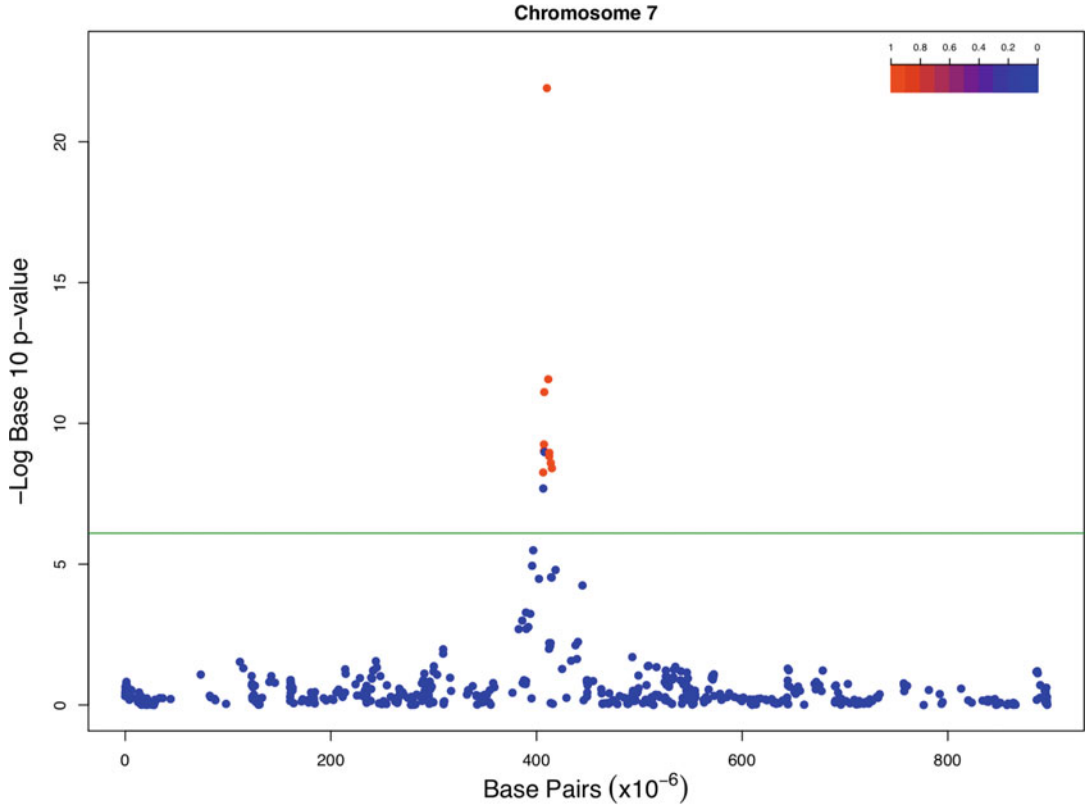

Fig. 3 Display of linkage disequilibrium on chromosome-wise Manhattan plot. The genome-wide association study was conducted on a simulated trait controlled by 20 genes with heritability of 0.75 in a mouse population with 1940 individuals genotyped with 12,000 SNPs. Data for chromosome 7 is shown to demonstrate the linkage disequilibrium (R square) between the most significant marker and its neighboring markers

Manhattan plots created by multiple-loci models, which reduce the markers to bins and only present the marker with the lowest *P*-value in each bin. Now there are several multistep GWAS methods (such as multiple loci mixed linear model, FarmCPU, and BLINK) using likelihood value or Bayesian information criteria to filter the most significant markers (named "pseudo QTNs"). The continuous peaks of NYC-type Manhattan plots are used to indicate existing large QTL. In the simulation study (Fig. 4), these multiple loci or multistep strategies have greater power than a simpler method, and for some real traits, they are also reported to produce more credible results [71, 72]. These Kansas-type Manhattan plots are also very useful for distinguishing two or more close QTLs.

5 Examining QQ Plots

The quantile–quantile (QQ) plot compares two probability distributions by plotting the quantile values against each other [73]. All points are always nondecreasing from bottom left to upper right. In GWAS, the QQ plot helps to identify the inflation of *P*-values and the markers exceeding the expectation. Under the null hypothesis, all *P*-values follow a uniform [0, 1] distribution. In GWAS, one expectation is that most of the markers do not associate with the trait and only a small proportion do. Therefore, most markers' dots will lie on the diagonal line in the QQ plot and some deviated markers are off the diagonals (Fig. 5).

5.1 Expectation

For comparison between the testing results and the null hypothesis, the *P*-value will be used to calculate the expectation of such *P*-values (EP-value). After sorting all *P*-values from 0 to 1, the P_k can be used to present $P_1, P_2, \ldots, P_n$ ($k = 1, 2, \ldots, n$) [74], where n is the total number of markers in the GWAS testing. The expected probability of all marker effects follows a uniform distribution, so the expectation P_k (EP_k) can be calculated with $k/(n + 1)$. Like *P*-values, EP-values are also converted to negative logarithms.

5.2 Confidence Interval

A confidence interval gives a range within which a statistic can be off the expectation [75]. In the GWAS, *P*-values follow a beta distribution under the null hypothesis that markers are not associated with the trait [76]. After log transformation, the confidence interval of small *P*-values is larger than the confidence interval of large *P*-values. Thus, in the QQ plot, the confidence interval area is changing from narrow left (large *P*-values) to wide right (small *P*-values).

5.3 Inflation and Deflation

The median of the observed *P*-values is expected to equal the median of expected *P*-values. If let λ stand for the ratio between the median of observed *P*-values and the median of expected *P*-

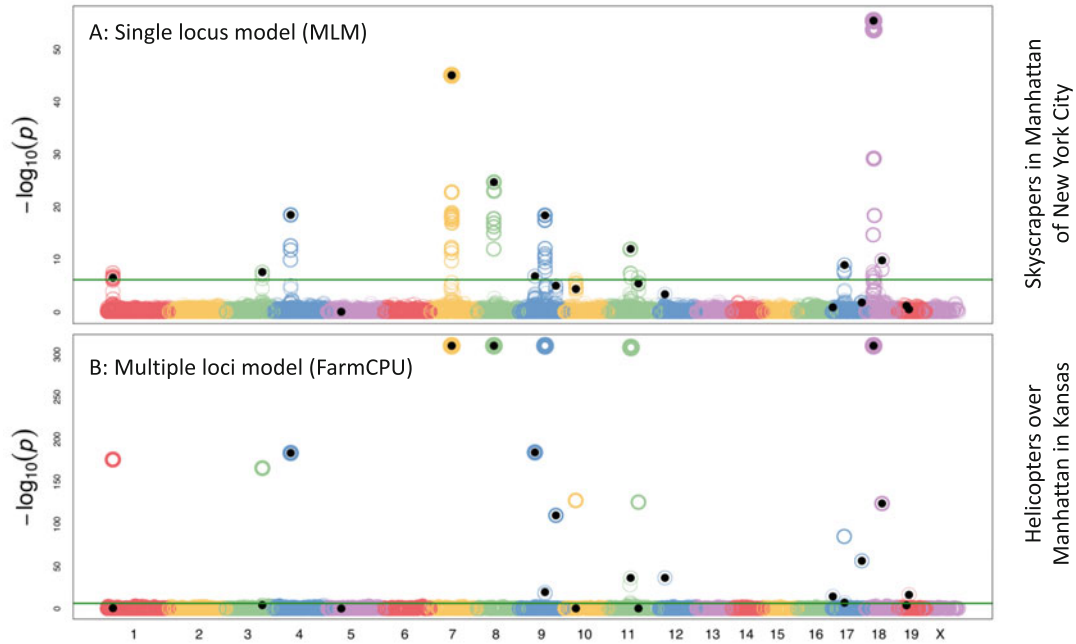

Fig. 4 Two types of Manhattan plots resulting from single-locus and multiple-loci models. The genome-wide association studies were conducted using a single-locus model (MLM) and multiple-loci model (FarmCPU) implemented in GAPIT. The population contains 1940 mice genotyped with 12,000 SNPs. The trait was simulated with 0.75 heritability and 20 genes. The associated markers in a linkage disequilibrium (LD) block appear as a spike (**a**). The spikes look like the skyscrapers in Manhattan of New York City. In the multiple-loci model, only one marker in an LD block can have a significant *P*-value (**b**). The scattered associated markers appear like helicopters flying over Manhattan in Kansas where there are no skyscrapers

values, two consequences could be observed: inflation ($\lambda > 1.1$) or deflation ($\lambda < 0.9$) [26, 77]. Inflation indicates that most tests are systematically more significant than the expected distribution. Inflation occurs with lack of control of population stratification and unknown family relationships. Deflation indicates that most of the points are systematically less significant than the expected distribution. A common cause of deflation is that markers are assumed to be independent from each other and actually they are not, which is common when a linkage mapping population is used for GWAS.

6 Other Outputs

Although Manhattan and QQ plots are the major graphs used to present GWAS results, genotype distribution, estimated genetic parameters, phenotype analysis, and population structure also provide necessary and complementary information for the interpretation about the data and results (Fig. 6).

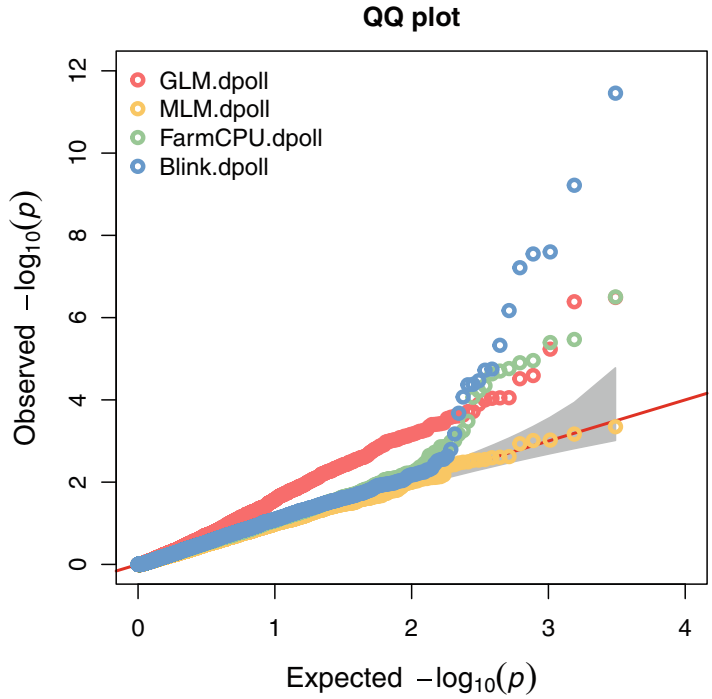

Fig. 5 Quantile-quantile (QQ) plot of genome-wide association study. The analysis was conducted using FarmCPU on 281 maize inbreds phenotyped for flowering time (days to pollination) and genotyped with 3093 SNPs. The shaded area indicates the 95% of confidence interval

6.1 Genotype Analysis

Rare SNPs with low MAF usually cause false positives, especially for small populations and when phenotypes do not follow a normal distribution [57, 78]. However, many causal genetic variants are rare [78], so the MAF distribution should be noticed. When MAF of markers are plotted against their *P*-values, extreme attention should be paid to the markers with small *P*-values and small MAF. The frequency of heterozygosity can be calculated for both individuals and markers. A high level of heterozygosity in a few inbred lines may indicate contamination during their development. A high level of heterozygosity across all inbred lines may suggest the problem of calling markers (Fig. 6a, b).

6.2 Distribution of Missing Genotypes

Similar to heterozygosity, missing genotypes can be analyzed in two directions, individual wise and marker wise. The missing genotypes can be imputed using special software, such as FILLIN [79], Impute [80] or Beagle [81]. The imputed markers should be prudently selected as candidate genes in a GWAS result, because their genotypes were imputed using other genotype information.

6.3 Kinship

The kinship matrix is visualized via a heat map (Fig. 6c). The *y*- and *x*-axes are the order of the individual taxa [31, 82]. The shade of

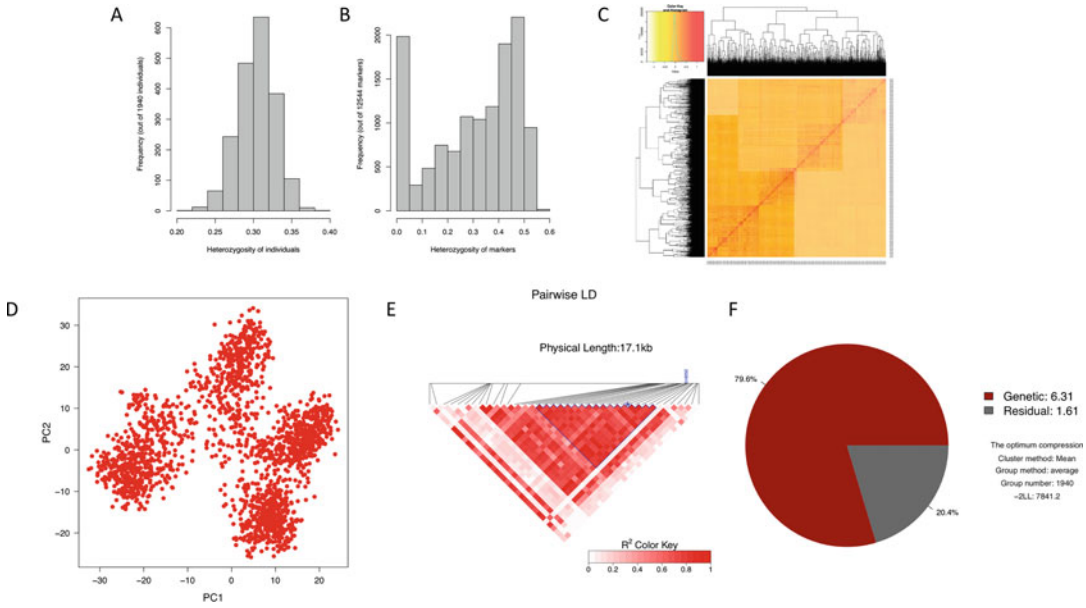

Fig. 6 The other output results in the GWAS. The demonstration data include 1940 individual mice and 12K markers. The genotype heterozygosity of individuals (**a**) and markers (**b**) are used to show the heterozygosity distribution. The kinship (**c**) and principal component (PC) (**d**) plots are used to reveal the relationship and population structure. (**e**) is the pairwise LD plot between the most significant marker and its neighboring markers. The heritability plot (**f**) is estimated by MLM

color in the squares indicates the relationship levels between each pair of individuals. Some software packages apply a distribution plot of kinship value, which helps users to evaluate the clustered relationships in the whole population.

6.4 Population Structure

Population structure and cryptic relatedness are important sources of spurious association. Population structure is usually fitted as a covariate in GWAS. The population structure can be quantified by two major approaches using genetic markers, including STRUCTURE [83] and principal component analysis (PCA) [84]. Different types of graphs are produced by these two methods. The Q matrix in the STRUCTURE [85] software is used to indicate the proportions of individuals belonging to different subpopulations. PCA uses a dimensionality reduction strategy to extract the eigenvalues and vectors from all genetic markers. Pairwise plots of principal components such as the first and the second principal component (Fig. 6d) are used to display the population's structure. The clustering trend indicates the relative population structure.

6.5 Linkage Disequilibrium Decay

Usually, LD is measured as the r^2 or D value for pairwise markers in a user-selected segment. The plot represents distances between two markers in the window and their squared correlation coefficient (Fig. 6e). Another presentation of LD is to plot LD between

markers against their distance to show the LD decay over distance. The moving average of adjacent markers is usually calculated by using sliding windows whose size can be set. The decreasing trend of the moving average windows shows the speed of LD decay in the population and is used to estimate the linkage distance. The LD decay also can be used to estimate the relative evolutionary relationship. Slow decay means more closely related species or individuals.

6.6 Heritability Estimation and Phenotype Distribution

Heritability is an important factor in genetic analyses, including GWAS. Low heritability means that only a small proportion of variability in the trait is explained by genetics and low statistical power is expected to conduct a GWAS [86, 87]. Therefore, it is necessary to understand the heritability level before performing GWAS. The ratio of genetic variance to total phenotypic variance is defined as the heritability (Fig. 6f). The phenotype distribution is another factor influencing statistical power for GWAS. Most GWAS methods assume that the residuals in the entire sample population follow a normal distribution. Illustration of phenotype distribution is helpful to identify data structure, outliers, and relationships between traits.

7 Interactive Outputs Using HTML

Compared to static outputs in formats such as PDF or JPG, interactive outputs using HTMP provide the opportunity for users to gain additional information. The information on a Manhattan plot (Fig. 7) includes minor allele frequency (MAF), estimated effect, names of neighboring gene, and the ratio of markers explaining genetic variance in the whole phenotypic variance. This information will help researchers more efficiently select candidate genes for downstream confirmatory experiments. An interactive Manhattan plot allows the user to select a chromosome, zoom in and out on the whole plot, or filter markers based on cutoff.

7.1 HTML

Hypertext Markup Language (HTML) is the most popular tool markup language designed for web browsers. HTML provides an approach to create structured multimedia documents that can contain images, text, and other interactive objects which are sub-elements in individual tags. Importantly, HTML can present information upon mouseover of the tags, making them more operative and interactive with users. Some R packages (such as lattice [88], rgl, scatterplot3, plot3d, and pca3d) have used this technology to present the 3D PCA plot, making it possible to rotate and zoom in on the whole figure to visualize the internal structure.

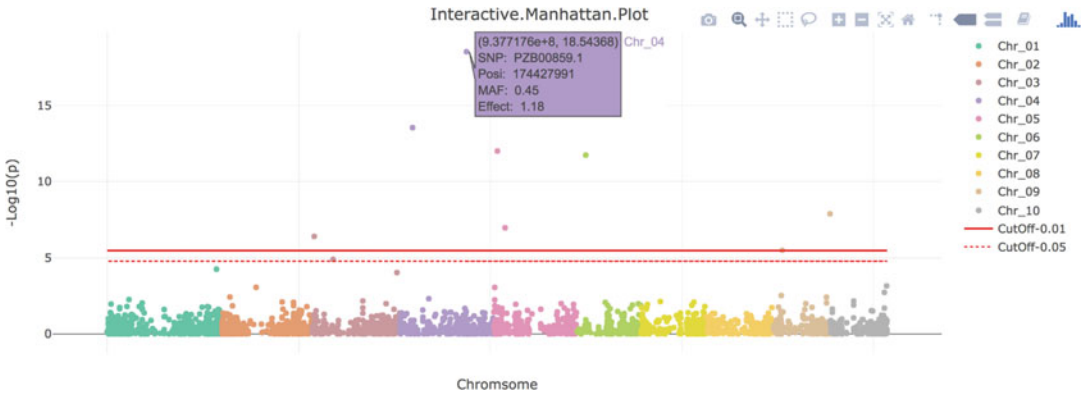

Fig. 7 The interactive Manhattan plot generated by GAPIT. The genome-wide association study was conducted on a trait simulated from 281 maize inbreds genotyped with 3093 SNPs. The trait was controlled by ten genes with heritability of 0.75. The plot is displayed using HTML format. When the cursor is near a point, the corresponding information is displayed for the SNP

7.2 R Library

The R package "plotly" applies an approach from R to HTML [89]. Each marker in the Manhattan plot is added a window with all information. When the mouse meets the marker, all information such as MAF, *P*-value, estimated effect, explained variance, and gene name are presented in the pop-up window. Each data type can occupy an individual row in an information block. The same type of data can be incorporated into a QQ plot. Using the plotly R package, these interactive Manhattan and QQ plots have been implemented in the GAPIT software. The plot can be displayed in web browsers, which requires a supporting folder named "library."

8 Final Remarks

Manhattan plots are the most common output to visually display the associations of genetic markers with traits of interest. Stacking multiple Manhattan plots in the circle or Cartesian format helps to demonstrate pleiotropy among multiple traits or the overlap among different models. QQ plots are essential to assess the quality and power of the GWAS by displaying the inflation/deflation of *P*-values and markers that exceeded the expectation. The additional information such as MAF, marker effect estimates, and detailed locations can be displayed as tabulate tables static graphic output, or interactive output of Manhattan and QQ plots using HTML. Finally, it is critical to visualize the properties of phenotypes and genotypes to identify the statistical power and source of spurious association. The properties include marker density, LD decay, MAF, phenotypic clusters and outliers, population structure, heritability, heterozygosity, and missing rate.

Acknowledgments

This project was partially funded by the National Science Foundation of the United States (Award # DBI 1661348 and ISO 2029933), the United States Department of Agriculture - National Institute of Food and Agriculture (Hatch project 1014919, Award #s 2018-70005-28792, 2019-67013-29171, and 2020-67021-32460), the Washington Grain Commission, the United States (Endowment and Award #s 126593 and 134574), the Program of Chinese National Beef Cattle and Yak Industrial Technology System, China (Award # CARS-37), Fundamental Research Funds for the Central Universities, China (Southwest Minzu University, Award # 2020NQN26), and Sichuan Science and Technology Program, China (Award #s 2021YJ0269 and 2021YJ0266).

Author Contributions Jiabo Wang: software, data curation, writing—original draft preparation, visualization, investigation.

Alexander E. Lipka: revision of the manuscript.

Jianming Yu: revision of the manuscript.

Zhiwu Zhang: conceptualization and revision of the manuscript.

References

1. Dimensionality M, Pan Q, Hu T et al (2013) Genome-wide association studies and genomic prediction, vol 1019. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-447-0
2. Yano K, Yamamoto E, Aya K et al (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet 48:927–934. https://doi.org/10.1038/ng.3596
3. Gibson G (2010) Hints of hidden heritability in GWAS. Nat Genet 42:558–560. https://doi.org/10.1038/ng0710-558
4. Lee C-Y, Kim T-S, Lee S et al (2015) Concept of genome-wide association studies. In: Current technologies in plant molecular breeding. Springer, New York, NY, pp 175–204. https://doi.org/10.1007/978-94-017-9996-6_6
5. Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. PLoS Comput Biol 8:e1002822. https://doi.org/10.1371/journal.pcbi.1002822
6. Wang MH, Cordell HJ, Van Steen K (2019) Statistical methods for genome-wide association studies. Semin Cancer Biol 55:53–60. https://doi.org/10.1016/j.semcancer.2018.04.008
7. Chen K, Baxter T, Muir WM et al (2007) Genetic resources, genome mapping and evolutionary genomics of the pig (Sus scrofa). Int J Biol Sci 3:153–165
8. Benson AK, Kelly SA, Legge R et al (2010) Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc Natl Acad Sci 107:18933–18938. https://doi.org/10.1073/pnas.1007028107
9. Andreescu C, Avendano S, Brown SR et al (2007) Linkage disequilibrium in related breeding lines of chickens. Genetics 177:2161–2169. https://doi.org/10.1534/genetics.107.082206
10. Zhu XM, Shao XY, Pei YH et al (2018) Genetic diversity and genome-wide association study of major ear quantitative traits using high-density SNPs in maize. Front Plant Sci 9:1–16. https://doi.org/10.3389/fpls.2018.00966
11. Zhang H, Yin L, Wang M et al (2019) Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front Genet 10:1–10. https://doi.org/10.3389/fgene.2019.00189

12. Pereira HD, Marcelo J, Viana S et al (2018) Relevance of genetic relationship in GWAS and genomic prediction Relevance of genetic relationship in GWAS and genomic prediction. J Appl Genet 59:1. https://doi.org/10.1007/s13353-017-0417-2
13. Stich B, Melchinger AE (2009) Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and Arabidopsis. BMC Genomics 10: 94. https://doi.org/10.1186/1471-2164-10-94. 1471-2164-10-94 [pii]
14. Benjamini Y, Yekutieli D (2005) Quantitative trait loci analysis using the false discovery rate. Genetics 171:783–790. https://doi.org/10.1534/genetics.104.036699
15. Schork AJ, Thompson WK, Pham P et al (2013) All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet 9:e1003449. https://doi.org/10.1371/journal.pgen.1003449
16. Yan G, Qiao R, Zhang F et al (2017) Imputation-based whole-genome sequence association study rediscovered the missing QTL for lumbar number in Sutai pigs. Sci Rep 47:615. https://doi.org/10.1038/s41598-017-00729-0
17. Zhang Z, Ober U, Erbe M et al (2014) Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS One 9: e0093017. https://doi.org/10.1371/journal.pone.0093017
18. Han Y, Zhao X, Cao G et al (2015) Genetic characteristics of soybean resistance to HG type 0 and HG type 1.2.3.5.7 of the cyst nematode analyzed by genome-wide association mapping. BMC Genomics 16:1–11. https://doi.org/10.1186/s12864-015-1800-1
19. Sukumaran S, Reynolds MP, Sansaloni CP (2018) Genome-wide association analyses identify QTL hotspots for yield and component traits in durum wheat grown under yield potential, drought, and heat stress environments. Front Plant Sci 9:81. https://doi.org/10.3389/fpls.2018.00081
20. Martinez SA, Godoy J, Huang M et al (2018) Genome-wide association mapping for tolerance to preharvest sprouting and low falling numbers in wheat. Front Plant Sci 9:1–16. https://doi.org/10.3389/fpls.2018.00141
21. Saatchi M, Schnabel RD, Taylor JF et al (2014) Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC Genomics 15:442. https://doi.org/10.1186/1471-2164-15-442
22. Tan B, Ingvarsson PK (2019) Integrating genome-wide association mapping of additive and dominance genetic effects to improve genomic prediction accuracy in Eucalyptus. BioRxiv 2019:841049. https://doi.org/10.1101/841049
23. Wei X, Zhang J (2016) The genomic architecture of interactions between natural genetic polymorphisms and environments in yeast growth. Genetics 205:genetics.116.195487. https://doi.org/10.1534/genetics.116.195487
24. Vinkhuyzen AAE, Pedersen NL, Yang J et al (2012) Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. Transl Psychiatry 2: e125. https://doi.org/10.1038/tp.2012.49
25. Chen CY, Misztal I, Aguilar I et al (2011) Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: an example using broiler chickens. J Anim Sci 89:23–28. https://doi.org/10.2527/jas.2010-3071
26. Gusev A, Bhatia G, Zaitlen N et al (2013) Quantifying missing heritability at known GWAS loci. PLoS Genet 9:e1003993. https://doi.org/10.1371/journal.pgen.1003993
27. Eichler EE, Flint J, Gibson G et al (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450. https://doi.org/10.1038/nrg2809
28. Bradbury PJ, Zhang Z, Kroon DE et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. https://doi.org/10.1093/bioinformatics/btm308
29. Yang J, Lee SH, Goddard ME et al (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011
30. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575
31. Tang Y, Liu X, Wang J et al (2016) GAPIT Version 2: an enhanced integrated tool for genomic association and prediction. Plant Genome 9(2):1. https://doi.org/10.3835/plantgenome2015.11.0120
32. Kaur S, Zhang X, Mohan A et al (2017) Genome-wide association study reveals novel genes associated with culm cellulose content in bread wheat (Triticum aestivum, L.). Front Plant Sci 8:1–7. https://doi.org/10.3389/fpls.2017.01913

33. Hickey JM (2013) Genome-wide association studies and genomic prediction, vol 1019. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-447-0
34. Hayes B (2013) Genome-wide association studies and genomic prediction, vol 1019. Humana Press, Totowa, NJ, pp 149–169. https://doi.org/10.1007/978-1-62703-447-0
35. Ziegler A, König IR, Thompson JR (2008) Biostatistical aspects of genome-wide association studies. Biom J 50:8. https://doi.org/10.1002/bimj.200710398
36. Almli LM, Duncan R, Feng H et al (2014) Correcting systematic inflation in genetic association tests that consider interaction effects application to a genome-wide association study of posttraumatic stress disorder. JAMA Psychiatry 71:1392–1399. https://doi.org/10.1001/jamapsychiatry.2014.1339
37. Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155–160. https://doi.org/10.1016/j.copbio.2006.02.003
38. Gianola D, De Los CG, Hill WG et al (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363. https://doi.org/10.1534/genetics.109.103952
39. Evangelou E, Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14: 379–389. https://doi.org/10.1038/nrg3472
40. González-Camacho JM, de Los CG, Pérez P et al (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet 125:759–771. https://doi.org/10.1007/s00122-012-1868-9
41. Huang M, Liu X, Zhou Y et al (2018) BLINK : a package for the next level of genome-wide association studies with both individuals and markers Meng Huang. Gigascience 8:1–12. https://doi.org/10.1093/gigascience/giy154
42. Liu X, Huang M, Fan B et al (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12:e1005767. https://doi.org/10.1371/journal.pgen.1005767
43. Segura V, Vilhjálmsson BJ, Platt A et al (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44: 825–830. https://doi.org/10.1038/ng.2314
44. Kammerer S, Roth RB, Reneland R et al (2004) Large-scale association study identifies ICAM gene region as breast and prostate cancer susceptibility locus. Cancer Res 64: 8906–8910. https://doi.org/10.1158/0008-5472.CAN-04-1788
45. Yin L, Zhang H, Tang Z et al (2020) rMVP: a memory-efficient, visualization-enhanced, and parallel-1 accelerated tool for genome-wide association study. BioRxiv
46. Turner S (2018) qqman: an R package for visualizing GWAS results using Q-Q and Manhattan plots. J Open Source Softw 3:371. https://doi.org/10.1101/005165
47. Barrett JC, Fry B, Maller J et al (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263. https://doi.org/10.1093/bioinformatics/bth457
48. Brown PJ, Upadyayula N, Mahone GS et al (2011) Distinct genetic architectures for male and female inflorescence traits of maize. PLoS Genet 7:e1002383. https://doi.org/10.1371/journal.pgen.1002383
49. Ma J, Iannuccelli N, Duan Y et al (2010) Recombinational landscape of porcine X chromosome and individual variation in female meiotic recombination associated with haplotypes of Chinese pigs. BMC Genomics 11:159. https://doi.org/10.1186/1471-2164-11-159
50. Kover PX, Valdar W, Trakalo J et al (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. PLoS Genet 5:e1000551. https://doi.org/10.1371/journal.pgen.1000551
51. Buckler ES, Holland JB, Bradbury PJ et al (2009) The genetic architecture of maize flowering time. Science 325:714–718. https://doi.org/10.1126/science.1174276
52. Tian F, Bradbury PJ, Brown PJ et al (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43:159–162. https://doi.org/10.1038/ng.746
53. Lipka AE, Tian F, Wang Q et al (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399. https://doi.org/10.1093/bioinformatics/bts444. bts444 [pii]
54. Piepho HP (2001) A quick method for computing approximate thresholds for quantitative trait loci detection. Genetics 157:425–432
55. Connolly S, Heron EA (2014) Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits. Brief Bioinform 16:429–448. https://doi.org/10.1093/bib/bbu017

56. Churchill GA, Doerge RW (2008) Naive application of permutation testing leads to inflated type I error rates. Genetics 178:609–610. https://doi.org/10.1534/genetics.107.074609
57. de Bakker PIW, Yelensky R, Péer I et al (2005) Efficiency and power in genetic association studies. Nat Genet 37:1217–1223. https://doi.org/10.1038/ng1669
58. Ganjgahi H, Winkler AM, Glahn DC et al (2018) Fast and powerful genome wide association of dense genetic data with high dimensional imaging phenotypes. Nat Commun 9:3254. https://doi.org/10.1038/s41467-018-05444-6
59. Chen CW, Yang HC (2019) OPATs: omnibus P-value association tests. Brief Bioinform 20:1–14. https://doi.org/10.1093/bib/bbx068
60. Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilità. Pubbl Del R Ist Super Di Sci Econ e Commer Di Firenze
61. Dunn OJ (1961) Multiple comparisons among means. J Am Stat Assoc 56:52. https://doi.org/10.2307/2282330
62. Ingvordsen CH, Backes G, Lyngkjær MF et al (2015) Genome-wide association study of production and stability traits in barley cultivated under future climate scenarios. Mol Breed 35:84. https://doi.org/10.1007/s11032-015-0283-8
63. Simes RJ (1986) A improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754
64. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300. https://doi.org/10.2307/2346101
65. Zhao F, McParland S, Kearney F et al (2015) Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genet Sel Evol 47:49. https://doi.org/10.1186/s12711-015-0127-3
66. Doerge RW, Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. Genetics 142:285–294. https://doi.org/10.1111/j.1369-7625.2010.00632.x
67. Phipson B, Smyth GK (2010) Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. Stat Appl Genet Mol Biol 9:39. https://doi.org/10.2202/1544-6115.1585
68. Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4:587–597. https://doi.org/10.1038/nrg1123
69. De La Vega FM, Isaac H, Collins A et al (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. Genome Res 15:454–462. https://doi.org/10.1101/gr.3241705
70. Lipka AE, Kandianis CB, Hudson ME et al (2015) From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Curr Opin Plant Biol 24:110–118. https://doi.org/10.1016/j.pbi.2015.02.010
71. Jernigan KL, Godoy JV, Huang M et al (2018) Genetic dissection of end-use quality traits in adapted soft white winter wheat. Front Plant Sci 9:271. https://doi.org/10.3389/fpls.2018.00271
72. Hu G, Li Z, Lu Y et al (2017) Genome-wide association study identified multiple genetic loci on chilling resistance during germination in maize. Sci Rep 7:1–11. https://doi.org/10.1038/s41598-017-11318-6
73. Pleil JD (2016) QQ-plots for assessing distributions of biomarker measurements and generating defensible summary statistics. J Breath Res 10:035001. https://doi.org/10.1088/1752-7155/10/3/035001
74. Wilk MB, Gnanadesikan R (1968) Probability plotting methods for the analysis of data. Biometrika 55:1. https://doi.org/10.1093/biomet/55.1.1
75. Neyman J (1937) Outline of a theory of statistical estimation based on the classical theory of probability. Phil Trans R Soc London Ser A Math Phys Sci 236:333. https://doi.org/10.1098/rsta.1937.0005
76. Robinson GK (1975) Some counterexamples to the theory of confidence intervals. Biometrika 62:155. https://doi.org/10.2307/2334498
77. Holland D, Fan CC, Frei O et al (2017) Estimating inflation in GWAS summary statistics due to variance distortion from cryptic relatedness. BioRxiv. https://doi.org/10.1101/164939
78. Lee S, Abecasis GR, Boehnke M et al (2014) Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet 95:5. https://doi.org/10.1016/j.ajhg.2014.06.009
79. Swarts K, Li H, Romero Navarro JA et al (2014) Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7:1–12. https://doi.org/10.3835/plantgenome2014.05.0023

80. Howie B, Fuchsberger C, Stephens M et al (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44:955–959. https://doi.org/10.1038/ng.2354
81. Ayres DL, Darling A, Zwickl DJ et al (2012) BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst Biol 61:170. https://doi.org/10.1093/sysbio/syr100
82. Zhang Z, Buckler ES, Casstevens TM et al (2009) Software engineering the mixed model for genome-wide association studies on large samples. Brief Bioinform 10:664–675. https://doi.org/10.1093/bib/bbp050
83. Raj A, Stephens M, Pritchard JK (2014) FastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197:573. https://doi.org/10.1534/genetics.114.164350
84. Duan F, Ogden D, Xu L et al (2013) Principal component analysis of canine hip dysplasia phenotypes and their statistical power for genome-wide association mapping. J Appl Stat 40:235–251. https://doi.org/10.1080/02664763.2012.740617
85. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959. https://doi.org/10.1111/j.1471-8286.2007.01758.x
86. Saatchi M, Miraei-Ashtiani SR, Nejati Javaremi A et al (2010) The impact of information quantity and strength of relationship between training set and validation set on accuracy of genomic estimated breeding values. African. J Biotechnol 9:438–442. https://doi.org/10.5897/AJB09.1024
87. Daetwyler HD, Pong-Wong R, Villanueva B et al (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031. https://doi.org/10.1534/genetics.110.116855. genetics.110.116855 [pii]
88. Cheshire J (2009) Lattice: multivariate data visualization with R. J Stat Softw Bk Rev 25(2). https://doi.org/10.1111/j.1467-985x.2009.00624_12.x
89. Carson S, Chris P, Toby H, et al (2016) plotly: create interactive web graphics via "plotly. js." R Packag Version