Genome-wide association studies in plants

Anderson, Robyn[1]

Edwards, David[1]

Batley, Jacqueline[1]

Bayer, Philipp Emanuel[1*]

Affiliations: 1. School of Biological Sciences and Institute of Agriculture, The University of Western Australia, Perth 6009, Australia

Corresponding author: Philipp Bayer, philipp.bayer@uwa.edu.au

Abstract

Cheap genome sequencing technology has made it possible to search for genomic variants called Single Nucleotide Polymorphisms (SNPs) for hundreds of individuals. Linking these genomic variants to phenotypes is the main goal in running genome-wide association studies (GWAS). Here we introduce common methods to call SNPs and approaches to quality-control the resulting dataset. We then introduce the major mathematical approaches to perform GWAS, show some software packages that implement these methods, and summarise a few common approaches to interpret the results.

Key words (five to ten): Genome-wide association studies (GWAS), Generalized linear models (GLMs), mixed linear models (MLM), Brassica, wheat

Key concepts

- GWAS is a powerful tool to associate genomic variants with phenotypes
- Quality control is core to any good GWAS
- Numerous powerful tools now exist to make running a GWAS straightforward
- Interpreting the output is still a challenge, especially in the presence of hidden confounding factors

Introduction

Recent advances in genomics technologies have led to an explosion in available data. It is now possible to phenotype and genotype populations of hundreds, or even thousands, of plants. This data can then be used to run genome-wide association studies (GWAS). GWAS can associate differences in allele frequencies with differences in measured phenotypes. This has led to advances in understanding the genetic basis of many plant traits, from flowering time in soybean (Zhang et al., 2015) to fruit phenotypes in peach (Cao et al., 2016) and disease resistance in canola (Raman et al., 2016). GWAS has been more successful in plants, with Single Nucleotide Polymorphisms (SNPs) explaining a much greater amount of phenotypic variation in plants than in humans (Brachi et al., 2011).

GWAS studies combine large datasets of around a hundred or up to thousand diverse individuals, either from breeding programs, collected wildtypes or seed banks. These large sample sizes are required to provide sufficient statistical power, as the effect size of a single SNP is usually very small. In this short review we will describe common techniques for collecting genotypic variation, provide an overview of the statistical basis of GWAS, and explain common approaches to quality control.

Main text

Commonly used types, advantages and drawbacks of SNP data

Single nucleotide polymorphisms (SNPs) are the most common form of molecular genetic marker for genotypic analysis. Here we present the most commonly used methods to call SNPs in plants, some of which also describe the production of genetic linkage maps based on genetic recombination and inheritance. GWAS software use SNP tables as input, and do not require linkage maps, however these linkage maps can be useful in subsequent analyses and in determining the parental source of useful SNPs where a sequenced genome assembly is not available. Each of the methods described below has its own drawbacks and benefits, as discussed in Scheben et al. (2017).

SNP Chip (Array)

A common form of SNP chip is an array of oligonucleotide probes fixed to a solid surface. The probes are designed based on genomic data to bind to previously characterised SNP loci. SNP chips are relatively cheap to run, allow for fast genotyping of the same SNPs for large sample sets, and the results are relatively easy to analyse. However, the SNP chip first has to be designed using known SNPs and the cost per array decreases with the number of samples, which may make them more expensive to use on lesser studied species where few assays are performed.

Whole Genome Resequencing (WGR) and SNP calling

Whole Genome Resequencing (WGR) is an effective way to survey genomic variation within a population, as it surveys the entire the genome, and lacks the inherent biases of reduced representation sequencing. WGR involves sequencing the genomes of interest at a relatively low coverage (<5x), trimming the reads and then aligning them to a reference genome to call SNPs. The SNPs are called using dedicated software such as samtools or freebayes. WGR can find all SNP variants and does not require prior information, genotyping at a much higher resolution than SNP chips.

Depending on genome size or heterozygosity, WGR can be prohibitively expensive for many research groups, and involves the handling of large amounts of data, requiring substantial computational resources and expertise. To reduce these drawbacks, reduced representation methods have been developed.

SkimGBS

Skim-based genotyping by sequencing (Skim GBS) involves sequencing of candidate genomes at a relatively low coverage (<5x) (Bayer et al., 2015). It is most commonly used to genotype populations where the parents' SNPs have been called using high coverage data while the progeny genotypes are inferred using low coverage sequencing data. It has mostly been applied in homozygous populations, where it provides high resolution without the cost of higher coverage WGR.

Reduced Representation Methods

Reduced representation sequencing methods aim to decrease the complexity of the sample DNA before sequencing. It can involve techniques from exome sequencing to restriction-site associated DNA sequencing (RADseq). The smaller representation of the genome reduces sequencing cost, but the smaller amount of data produced means that there will be fewer SNPs called, and the SNPs called may differ between individuals and experiments. Due to the low cost, reduced representation-based methods have become the tool of choice in ecological studies where there often is no reference genome or SNP chip (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016).

Exome capture

First developed in humans Exome capture is a technology that uses known coding regions from genome sequencing projects, and designs probes similar to the above SNP microarray technology (Ng et al., 2009). Captured regions are then sequenced, which greatly saves on cost, but SNPs located in non-coding regions will be dropped.

RAD Seq

Restriction-site associated DNA Sequencing (RADseq) aims to reduce the amount of the genome sequenced through the selection of genomic regions next to restriction enzyme sites (Baird et al., 2008). In RADseq, genomic DNA is extracted, digested with a restriction enzyme and adapters are ligated to the fragments. These DNA fragments are then sheared via sonication to reduce their size, and a second adapter is ligated to the sheared ends of the genomic DNA. PCR is then used to selectively amplify fragments with one of each adapter, and these are subsequently sequenced.

The reads produced are trimmed, identical sequences are grouped, and then counted. Polymorphic RAD tags are defined, then mapped to the reference genome, first identifying exact matches, and then alignments with 1 mismatch. The position in the genome where this mismatch occurs is the SNP, and these are then examined in terms of inheritance. Different SNP calling software uses different mismatch cut-offs to reduce the number of false positive SNPs.

No prior genomic knowledge is required for RAD Seq, and the adapters are compatible with barcoding, and so allow multiplexing for efficient and cheaper sequencing. The number of SNPs for subsequent analysis can be influenced by the choice of restriction enzyme: the more frequently the enzyme cuts the DNA, the more fragments that will be produced and therefore more fragments are likely to be sequenced.

ddRAD Seq

Double digested restriction-site associated DNA (ddRAD) sequencing was developed from the original RADseq method, and involves the digestion of genomic DNA with two restriction enzymes (Peterson et al., 2012) rather than one enzyme and shearing. The second enzyme was added to minimise the bias introduced by using only one enzyme. Everything else remains the same as in RAD seq.

How to run a GWAS

Here we lay out the steps necessary to run a GWAS, from SNPs to final visualisations, with a focus on plant genetic data. The input for all GWAS programs consists of a table of SNP calls in a population, and a table of phenotypes.

Quality control

Quality control (QC) of the input data is the most important step in any GWAS, as unchecked errors can have large effects on the statistical power and results. There are several common steps that can be carried out with most GWAS software such as TASSEL or PLINK. Common filters include:

- Minor allele frequency (MAF). False positive SNPs due to sequencing or genotyping errors often have an alternative allele only in one individual, which leads to a low allele frequency. It is common to remove all SNPs with a MAF below 5%. However, especially in humans there are many rare causative variants (see discussion of several papers in Shields (2011)), so caution is advised.
- Missingness. It is common to remove individuals or SNPs with many missing alleles, as this lack of information indicates poor genotyping quality. However, with genotyping arrays, ‘missing’ allele input can be treated as presence/absence variation, in which case filtering for missingness removes important associations (Gabur et al., 2018). The exact percentage of missingness depends on the dataset.
- Hardy Weinberg Equilibrium (HWE). Under Hardy-Weinberg equilibrium it is assumed that genotype frequencies will occur at a specific ratio dependant on the allele frequencies within the population. SNPs that are not in HWE are commonly removed from GWAS analysis, however, both artificial and natural selection can lead to SNPs that are outside of HWE, and these filtered SNPs can be viable candidate SNPs.

- Population outliers. Identity-by-descent (IBD) neighbour analysis can be used to find any individual outliers which are distinct from all other members of the population. These outliers should be removed as they likely carry rare alleles, leading to overinflated estimates of their contribution to traits of interest.

In polyploid crop species additional filtering may be necessary due to the presence of multiple genomes that may be highly similar to each other. Polyploid genetic data that is not handled appropriately can introduce false heterozygosity, or false-positive SNPs, due to mismapping of sequencing reads. Some filtering approaches remove all heterozygous SNPs under the assumption that many plant species are mostly homozygous (Lorenc et al., 2012). More sophisticated approaches, such as SWEEP, use subgenome specific haplotypes to remove false SNPs (Clevenger and Ozias-Akins, 2015).

Population stratification

A major problem with GWAS is that population stratification can lead to different allele frequencies and cause false positive associations. If members of population A share a certain allele more often than expected because all members inherit that allele from a shared ancestor, and if members of the same population exhibit a disease more often, if we do not account for population structure, this will lead to the allele becoming erroneously associated. Several methods have been proposed to correct for population stratification.

The most common method of correcting for stratification is principal components analysis (PCA), first proposed by (Price et al., 2006). PCA works by transforming a matrix of possibly correlated SNPs into a set of uncorrelated variables called principal components (PCs). Most PCA software calls 20 PCs by default, but most GWAS only use up to three PCs as covariates, and usually two PCs are enough to distinguish most populations. An example was published by Novembre et al. (2008), where the ancestry of 3,000 Europeans could be differentiated using two PCs.

If too few PCs are chosen, the model is not able to distinguish subpopulations (as happened in for example Peloso and Lunetta (2011)). In contrast, if too many PCs are chosen the model loses statistical power and no significant SNPs will be reported. As a first step, the first two or three PCs should be plotted against each other to see whether their groupings correlate with what is known about the individuals' ancestry. The variance explained by PCs can also be calculated in order to decide how many PCs are necessary, or GWAS can be run several times using one, two, three, or more PCs as covariates to investigate how the trait associated SNPs change.

Other approaches to population modelling based on SNPs exist, and the most well-known approach based on population structure modelling is STRUCTURE (Pritchard et al., 2000), which was followed by fastSTRUCTURE (Raj et al., 2014). STRUCTURE, and similar software packages, can visualise population structure to reconstruct the genetic history of populations, and the groups assigned by STRUCTURE have been shown to correlate with PCA-based assignments (Ma and Amos, 2012). Despite this ability to investigate population structure and its correlation with PCA-based assignments, one must be careful not to misinterpret the resulting plots (Lawson et al., 2018).

After QC and modelling of the population structure, it is finally possible to run the GWAS.

Finding associated SNPs

All GWAS approaches calculate and assign an effect size and a p-value to the association of a SNP with a phenotype. The effect size is measured using different values depending on the method used (for example, odds ratio) and quantifies the size of the difference between two groups. The associated p-value is an often misunderstood measurement. Many researchers think that it represents the *strength* of the association, which is not necessarily true. The p-value measures the probability of seeing the SNP association's observed summary statistic (odds ratio etc.) if no such association actually exists. A p-value of 0.05 therefore means that there is a five percent chance that the observed effect size describing the relationship between the SNP and the trait is calculated if there is no association in reality.

In GWAS, several different mathematical approaches are used to calculate effect size and the associated p-values. By themselves these approaches do not account for population structure or other confounding sources.

Generalized linear models (GLMs) and mixed linear models (MLMs) are commonly used in plant GWAS and both are flexible and powerful variants of regressions. Generalized linear models (GLMs) are a group of models including multiple regressions and ANOVA which do not account for population structure. Mixed linear models (MLMs) were introduced later as an improved approach to account for population structure and cryptic relationships between individuals (Yu et al., 2006). MLMs account for cryptic relationships by calculating a kinship matrix which shows which individuals are more closely related to each other. Most modern GWAS software implements GLMs, MLMs, and other approaches based on both, letting the user run both to decide which SNPs are most important.

After the GWAS software has finished, the first area of interest is usually the p-values reported for the association of SNPs with phenotypes of interest. We are interested in p-values below a certain cut-off, and often this cut-off for statistical significance is 0.05. The more tests are run, the more it is expected to see SNPs that randomly fall below that p-value cut-off. In GWAS, one test is run per SNP, so if one million SNPs are tested for association with a p-value cut-off of 0.05, we expect 50,000 SNPs to fall below the cut-off just by chance (1,000,000 * 0.05 = 50,000). In the past, a lower significance cut-off has been proposed, such as $5 * 10^{-8}$, as suggested in Fadista et al. (2016), which would lead to less than one false positive association (1,000,000 * $5e^{-8}$ = 0.05). Instead of altering the statistical significance cut-off, it is also possible to correct for multiple testing.

There are several different methods to correct for multiple testing. When using the PLINK software package, specifying the --adjust flag will run the most common tests, such as Bonferroni correction or Genomic Control. In the presence of linkage disequilibrium these values can be too conservative, and additional permutation tests are necessary (Gao et al., 2010). Alternatively R-packages, such as qvalue, can be used which can intake unadjusted p-values, and calculate corrected q-values (Dabney et al., 2010).

Validation of results

There are several tools to use to sanity-check the final results of the GWAS. The most popular one is the Q-Q (quantile-quantile) plot, which plots the obtained p-values against a theoretical distribution of p-values. In a perfect dataset, the majority of p-values follow the expected distribution, with a few strong outlier p-values which do not follow that distribution – these are our candidate alleles. However, in many cases, the line of p-values does not follow the theoretical distribution. One explanation for that is

the wrong theoretical distribution (see for example Voorman et al. (2011)). More common is a missing adjustment for population structure. In that case, the line of p-values will not follow the line of theoretical p-values.

A common visualisation of GWAS results is the Manhattan plot, which was named after their resemblance to a skyline of a large city. On the x-axis are the chromosomes and their positions, on the y-axis is the negative log of the p-values ranging from 0 to roughly 10, where each dot corresponds to one SNP's p-value. In a well-adjusted dataset, only few outliers appear, and these are the candidate SNPs (Figure 1). In a dataset with strong population stratification, skewed phenotypes, or insufficient quality control, many more p-values will appear above the cut off line (Figure 2).

Figure 1: Example of a well-adjusted Manhattan plot, with a few identified SNPs in red above the cut off line

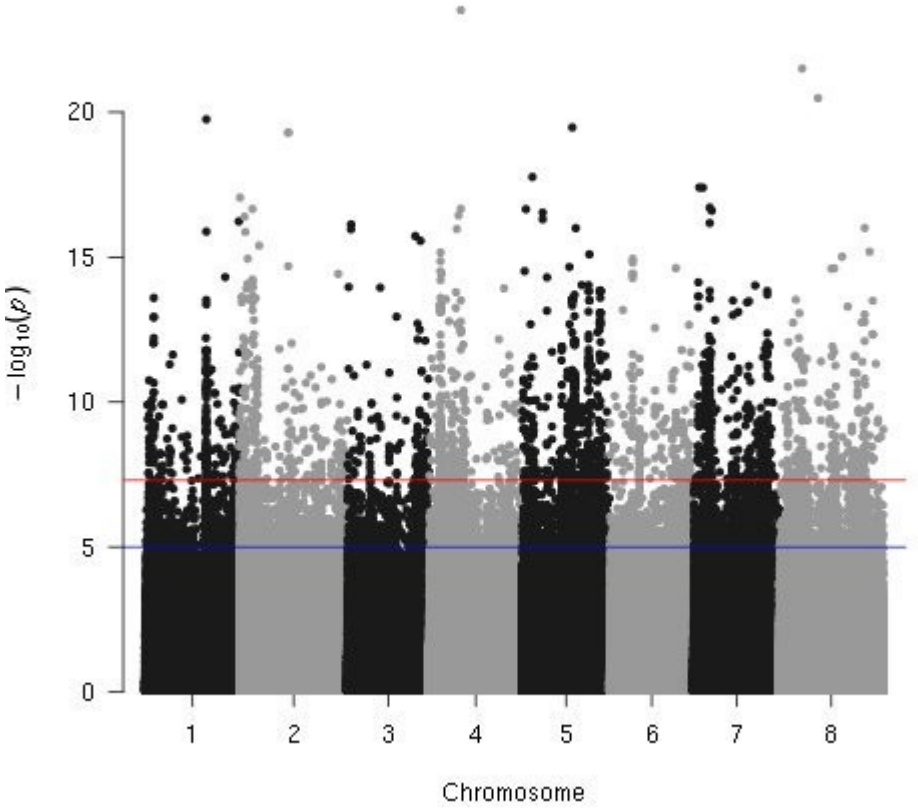

Figure 2: Example of a GWAS where the majority of SNPs are spuriously linked with the phenotype.

Depending on the SNP-calling method, the density of SNPs in the Manhattan plot will vary widely. With high density SNPs, the SNPs around the associated SNP may exhibit a 'hill-like' pattern where the closer SNPs are to an associated SNP, the lower their p-values are. With low density SNPs, low diversity, small population size, or recombinations this 'hill-like' pattern will not occur.

The best quality control is replication, ideally using an independent dataset, or using a different genotyping technology. Replication using a different technology will avoid GWAS false positives due to technological issues. Replication using a different population will avoid incorrect SNP identification due to population issues.

Software approaches

There are several commonly used software packages. Here we will give a short overview of the software usually used in plant GWAS (summarised in Table 1).

The main workhorse of GWAS is PLINK which combines a wide range of QC and data management functions with population stratification and association analyses. It can perform linear or logistic regressions with covariates, including methods to correct for multiple testing, but does not contain methods for more complex regressions such as MLM or GLM. The original PLINK 1 (Purcell et al., 2007) is

no longer under development, but a complete rewrite in PLINK 2 (Chang et al., 2015) is still being used and expanded.

PLINK was originally intended for use with human genotyping data which leads to some peculiarities when used with plant data. For example, PLINK expects that all chromosome names follow the human genome, a check which can, and in plants should, be turned off using the --allow-extra-chr flag. PLINK also expects the presence of X or Y chromosomes, which can be turned off using the --allow-no-sex flag.

TASSEL is one of the most common software packages used in GWAS (Bradbury et al., 2007), and it has both a graphical user and command line interface. It is important to note that TASSEL by default only reports unadjusted p-values, for which further corrections for multiple testing are necessary. Users searching for online tutorials on how to use TASSEL should be aware that TASSEL has had several major versions and that each version had a slightly different user interface, which can make older tutorials confusing.

GAPIT is an R-package that combines GWAS and genomic prediction methods (Lipka et al., 2012; Tang et al., 2016). GAPIT implements a single command which, given a set of SNPs, runs an entire GWAS starting with quality control to the plotting of the most common figures (QQ-plot, Manhattan plot etc.) and correcting for multiple testing. Using the parameter model.selection=TRUE it is possible to select the optimal number of principal components per phenotype. GAPIT supports running several models at once: GLM, MLM, CMLM, SUPER, FarmCPU, Blink, and MLMM. It is recommended to run several models as each has its strengths and drawbacks.

MVP (https://github.com/XiaoleiLiuBio/MVP) is a yet unpublished R-package which, similar to GAPIT, implements a single command which runs an entire GWAS. MVP implements three methods (GLM, MLM, FarmCPU) and like GAPIT, automatically draws Manhattan, Q-Q, and other plots. For large datasets MVP can be run in low memory mode.

Table 1: Overview of common GWAS software packages

Software	Citation	Approaches implemented	Data management functions	Plotting functionality	License
PLINK	(Chang et al., 2015; Purcell et al., 2007)	MLM	Yes	No	GPLv3
TASSEL	(Bradbury et al., 2007)	MLM, GLM	Yes	Yes	LGPLv2.1
GAPIT	(Lipka et al., 2012; Tang et al., 2016)	MLM, GLM, CMLM, FaST, SUPER, DC, FarmCPU, Blink, MLMM	No	Yes	?
MVP	https://github.com/XiaoleiLiuBio/MVP	MLM, GLM, FarmCPU	No	Yes	Apache License 2.0

In humans with large, and well-controlled datasets of hundreds of GWA-studies, it has been possible to perform meta-analyses, which take the results of each GWAS and calculate summary statistics to

remove spurious SNP/phenotype correlations and to find candidate SNPs shared by several studies. In plants there is a distinct lack of large central repositories storing GWAS results, but some first steps have been made in easyGWAS where the results of 313 GWA-studies are stored (Grimm et al., 2017).

Once candidate SNPs have been identified, they can be mapped to candidate genes based on gene and SNP position overlaps. This can be done manually, or automatically using tools like bedtools intersect/window (Quinlan & Hall, 2010). Many candidate SNPs are not within genes, but up- or downstream in regulatory regions of genes. Therefore it is important to not only search for direct SNP/gene overlaps, but to search within a 10 kb or 100 kb window around genes.

Future of GWAS in plants

What does the future hold for GWAS in plants? We now have enough data to skip the SNP calling step and go directly to k-mer based association studies. HAWK (Rahman et al., 2018), SEER (Lees et al., 2016) and pyseer (Lees et al., 2018) are tools which run genome-wide association studies with k-mers instead of SNPs. K-mers are obtained by breaking the sequencing reads themselves into pieces of length *k* (usually somewhere around 31). The k-mers are counted and differences in k-mer counts between case and control individuals are modelled, and k-mers that deviate from the expected distribution are treated as candidate k-mers. From there it is possible to map those k-mers back to a reference genome in order to learn which genes or regulatory regions they are associated with.

In plants, k-mer based association studies have been used successfully in wheat to find contigs containing different resistance gene candidates (Arora et al., 2018). No other application of k-mer association studies has been recorded yet in plants. The amount of data is limiting, but this will be less of an issue as sequencing costs continue to fall. We believe that k-mer association studies allow researchers to look at all of the genome without bias of the SNP caller or read alignment mechanism. However, open questions remain, such as how to account for population structure when performing k-mer based association studies. It is important to note that both GWAS and k-mer based association studies cannot find causative associations, they can only find SNPs or k-mers *associated* with phenotypes. Biological reasoning and further experiments are necessary to distinguish statistical association from true causation.

Further reading list

Voorman, A., Lumley, T., McKnight, B. and Rice, K. (2011) Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PloS one* **6**, e19416. – **A very useful explanation of Q-Q plots**

Jeff Barrett, How to read a genome-wide assocation study, http://genomesunzipped.org/2010/07/how-to-read-a-genome-wide-association-study.php - **A tutorial on reading GWAS results for the uninitiated.**

Lawson, D.J., van Dorp, L. and Falush, D. (2018) A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature communications* **9**, 3258. – **It is easy to misinterpret STRUCTURE plots, this paper hopefully helps to avoid that**

Duke Pauli, Use of GAPIT for Genome Wide Association Studies, http://pbgworks.org/sites/pbgworks.org/files/GAPIT_with_SYslides.pdf - **A tutorial on how to run GAPIT and interpret its output**

https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual - **The TASSEL 5 user manual**

Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol 8(12): e1002822. – **A much more in-depth overview of concepts behind GWAS with a focus on humans.**

Scheben, A., Batley, J. and Edwards, D. (2017) Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant biotechnology journal* **15**, 149-161. – **Very useful discussion on when to use which genotyping technology in plants**

Glossary

GWAS – Genome-wide association study, a genome-wide study which tries to link allelic variation to phenotypic variation

PCA – Principal component analysis, takes a set of correlated observations (here: SNPs) and transforms them into a smaller set of uncorrelated observations (principal components, PCs)

SNP – Single Nucleotide Polymorphism, a single base genetic variation which can be linked to phenotypic variation

GLM – General Linear Model, a group of models ranging from multiple regressions to ANOVA which do not account for population structure

MLM – Mixed Linear Model, an improvement on GLMs which calculates or uses a Kinship matrix to account for population structure and cryptic relationships

MAF – Minor Allele Frequency, the percentage of the minor allele in a SNP, can range from 0 to 0.5. SNPs with a very low MAF can be false positives.

References

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews: Genetics*, *17*(2), 81-92. doi:10.1038/nrg.2015.28

Arora, S., Steuernagel, B., Chandramohan, S., Long, Y., Matny, O., Johnson, R., Enk, J., Periyannan, S., Hatta, M.A.M. and Athiyannan, N. (2018) Resistance gene discovery and cloning by sequence capture and association genetics. *bioRxiv*, 248146.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one* **3**, e3376.

Bayer, P.E., Ruperao, P., Mason, A.S., Stiller, J., Chan, C.K., Hayashi, S., Long, Y., Meng, J., Sutton, T., Visendi, P., Varshney, R.K., Batley, J. and Edwards, D. (2015) High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in Cicer arietinum and Brassica napus. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* **128**, 1039-1047.

Brachi, B., Morris, G.P. and Borevitz, J.O. (2011) Genome-wide association studies in plants: the missing heritability is in the field. *Genome biology* **12**, 232.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633-2635.

Cao, K., Zhou, Z., Wang, Q., Guo, J., Zhao, P., Zhu, G., Fang, W., Chen, C., Wang, X., Wang, X., Tian, Z. and Wang, L. (2016) Genome-wide association study of 12 agronomic traits in peach. *Nature communications* **7**, 13246.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7.

Clevenger, J.P. and Ozias-Akins, P. (2015) SWEEP: A Tool for Filtering High-Quality SNPs in Polyploid Crops. *G3* **5**, 1797-1803.

Dabney, A., Storey, J.D. and Warnes, G. (2010) qvalue: Q-value estimation for false discovery rate control. *R package version* **1**.

Fadista, J., Manning, A.K., Florez, J.C. and Groop, L. (2016) The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet* **24**, 1202-1205.

Gabur, I., Chawla, H.S., Liu, X., Kumar, V., Faure, S., von Tiedemann, A., Jestin, C., Dryzska, E., Volkmann, S., Breuer, F., Delourme, R., Snowdon, R. and Obermeier, C. (2018) Finding invisible quantitative trait loci with missing data. *Plant biotechnology journal*.

Gao, X., Becker, L.C., Becker, D.M., Starmer, J.D. and Province, M.A. (2010) Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet. Epidemiol.* **34**, 100-105.

Grimm, D.G., Roqueiro, D., Salome, P.A., Kleeberger, S., Greshake, B., Zhu, W., Liu, C., Lippert, C., Stegle, O., Scholkopf, B., Weigel, D. and Borgwardt, K.M. (2017) easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies. *The Plant cell* **29**, 5-19.

Lawson, D.J., van Dorp, L. and Falush, D. (2018) A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature communications* **9**, 3258.

Lees, J., Galardini, M., Bentley, S.D., Weiser, J.N. and Corander, J. (2018) pyseer: a comprehensive tool for microbial pangenome-wide association studies. *bioRxiv*, 266312.

Lees, J.A., Vehkala, M., Valimaki, N., Harris, S.R., Chewapreecha, C., Croucher, N.J., Marttinen, P., Davies, M.R., Steer, A.C., Tong, S.Y., Honkela, A., Parkhill, J., Bentley, S.D. and Corander, J. (2016) Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature communications* **7**, 12797.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., . . . Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, *28*(18), 2397-2399. doi:10.1093/bioinformatics/bts444

Lorenc, M.T., Hayashi, S., Stiller, J., Lee, H., Manoli, S., Ruperao, P., Visendi, P., Berkman, P.J., Lai, K., Batley, J. and Edwards, D. (2012) Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology* **1**, 370-382.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., . . . Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, *461*(7261), 272-276. doi:10.1038/nature08250

Ma, J. and Amos, C.I. (2012) Principal components analysis of population admixture. *PloS one* **7**, e40115.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M. and Bustamante, C.D. (2008) Genes mirror geography within Europe. *Nature* **456**, 98-101.

Peloso, G.M. and Lunetta, K.L. (2011) Choice of population structure informative principal components for adjustment in a case-control study. *BMC Genet.* **12**, 64.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. and Hoekstra, H.E. (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one* **7**, e37135.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. and Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842. doi:10.1093/bioinformatics/btq033

Rahman, A., Hallgrímsdóttir, I., Eisen, M. and Pachter, L. (2018) Association mapping from sequencing reads using k-mers. *eLife* **7**, e32920.

Raj, A., Stephens, M. and Pritchard, J.K. (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573-589.

Raman, H., Raman, R., Coombes, N., Song, J., Diffey, S., Kilian, A., Lindbeck, K., Barbulescu, D.M., Batley, J., Edwards, D., Salisbury, P.A. and Marcroft, S. (2016) Genome-wide Association Study Identifies New Loci for Resistance to Leptosphaeria maculans in Canola. *Frontiers in plant science* **7**, 1513.

Scheben, A., Batley, J. and Edwards, D. (2017) Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant biotechnology journal* **15**, 149-161.

Shields, R. (2011) Common disease: are causative alleles common or rare? *PLoS Biol.* **9**, e1001009.

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A.E., Buckler, E.S. and Zhang, Z. (2016) GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *Plant Genome* **9**.

Voorman, A., Lumley, T., McKnight, B. and Rice, K. (2011) Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PloS one* **6**, e19416.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. and Buckler, E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* **38**, 203-208.

Zhang, J., Song, Q., Cregan, P.B., Nelson, R.L., Wang, X., Wu, J. and Jiang, G.L. (2015) Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. *BMC genomics* **16**, 217.