Effects of input data quantity on genome-wide association studies (GWAS)

Yan Yan

Department of Computer Science,
University of Saskatchewan,
110 Science Place,
Saskatoon, SK, S7N 5C9 Canada
Email: yan.yan@usask.ca

Connor Burbridge

Global Institute for Food Security,
University of Saskatchewan,
110 Gymnasium Place,
Saskatoon, SK, S7N 0W9 Canada
Email: connor.burbridge@gifs.ca

Jinhong Shi

Department of Computer Science,
University of Saskatchewan,
110 Science Place,
Saskatoon, SK, S7N 5C9 Canada
Email: jinhong.shi@usask.ca

Juxin Liu

Department of Mathematics and Statistics,
University of Saskatchewan,
McLean Hall,
Saskatoon, SK, S7N 5E6 Canada
Email: liu@math.usask.ca

Anthony Kusalik*

Department of Computer Science,
University of Saskatchewan,
110 Science Place,
Saskatoon, SK, S7N 5C9 Canada
Email: kusalik@cs.usask.ca
*Corresponding author

Abstract: Many software packages have been developed for Genome-Wide Association Studies (GWAS) based on various statistical models. One key factor influencing the statistical reliability of GWAS is the amount of input data used. In this paper, we investigate how input data quantity influences output of four widely used GWAS programs, PLINK, TASSEL, GAPIT, and FaST-LMM, in the context of plant genomes and phenotypes. Both synthetic and real data are used. Evaluation is based on *p*- and *q*-values of output SNPs, and Kendall rank correlation between output SNP lists. Results show that for the same GWAS program, different *Arabidopsis thaliana* datasets demonstrate similar trends of rank correlation with varied input quantity, but differentiate on the numbers of SNPs passing a given *p*- or *q*-value threshold. We also show that variations in numbers of replicates influence the *p*-values of SNPs, but do not strongly affect the rank correlation.

Keywords: GWAS; genome-wide association study; *Arabidopsis thaliana*; plant phenomics; plant genomics; PLINK; TASSEL; GAPIT; FaST-LMM; statistical power; input data quantity; epistasis.

Reference to this paper should be made as follows: Yan, Y., Burbridge, C., Shi, J., Liu, J. and Kusalik, A. (2019) 'Effects of input data quantity on genome-wide association studies (GWAS)', *Int. J. Data Mining and Bioinformatics*, Vol. 22, No. 1, pp.19–43.

Biographical notes: Yan Yan received her PhD in Biomedical Engineering from the University of Saskatchewan. She is a currently Postdoctoral Fellow in the Department of Computer Science at the University of Saskatchewan. Before that, she was a Postdoctoral Fellow in the Department of Computer Science at the University of Western Ontario. Her research interests include proteomics with a focus on advanced methods for peptide sequencing, population genomics on large scale genome-wide data analysis, and machine learning algorithms in computational biology. Her current research project is on advanced algorithms for genome-wide data analysis in plant genomes.

Connor Burbridge recently completed his Undergraduate Honours education at the University of Saskatchewan. He is currently employed as a Research Technician for the Global Institute for Food Security with a strong focus on Bioinformatics and Data Processing. His past work has included genotype by sequencing pipelines, comparative analysis of genome-wide association programs, a methodology for comparison of *de novo* RNAseq assembly programs as well as differential gene expression analysis. His current research is focused on small genome assembly, microbial sequence classification and root image processing. He hopes to pursue further education in the future and participate in exciting interdisciplinary research opportunities.

Jinhong Shi received her PhD in Biomedical Engineering from the University of Saskatchewan. She is a Postdoctoral Fellow in the Department of Computer Science at the University of Saskatchewan. Her general research interests include data analytics, proteomics, bioinformatics and applied machine learning and deep learning. Currently, she is specifically interested in applying machine learning and deep learning to Single Nucleotide Polymorphism (SNP) data to identify significant SNPs that affect antimicrobial resistance in bacteria and plant phenotypes.

Juxin Liu received her PhD degree in Statistics from the University of British Columbia. She is a Professor in the Department of Mathematics and Statistics, University of Saskatchewan. Her current research interests include measurement error/misclassification modelling, missing data analysis, and advanced computational methods. She has established a strong publication

record in both statistical journals (such as *Biometrics*, *Statistics in Medicine*) and applied journals (such as *American Journal of Epidemiology*, *Ecology*).

Anthony Kusalik is a Professor in the University of Saskatchewan's Department of Computer Science. At the U of Saskatchewan, he is a Member of the Division of Biomedical Engineering, Associate Member of the School of Public Health, Director of the Bioinformatics Program, and Member of the Bioinformatics and Computational Biology Research Laboratory. He obtained his PhD degree in 1988 from the University of British Columbia and has been a faculty member at the U of Saskatchewan since December, 1985. His research interests range from logic programming to machine learning to computational biology. However, his primary research interests centre on bioinformatics and computational biology. He works extensively with collaborators in a broad spectrum of life and health sciences. He is an IEEE, IEEE CS, and ISCB Member.

This paper is a revised and expanded version of a paper entitled 'Comparing Four Genome-Wide Association Study (GWAS) Programs with Varied Input Data Quantity' presented at 'BIBM 2018 (2018 IEEE International Conference on Bioinformatics and Biomedicine)', 3–6 December 2018, Madrid, Spain.

1 Introduction

Genome-Wide Association Studies (GWAS) have served as primary methods for the past decade for identifying associations between genetic variants and traits or diseases (Hirschhorn and Daly, 2005; Bush and Moore, 2012). The most often used genetic variants are Single Nucleotide Polymorphisms (SNPs), which are changes of single DNA base-pairs. In plant breeding, GWAS have been widely used to link genotypes of plants to phenotypes of interest, and provide valuable insights into the mechanisms of causal SNPs (and related genes) linked to complex traits (Wang et al., 2012; Branham et al., 2015).

GWAS perform statistical hypothesis tests for each SNP, with the null hypothesis being no association between the SNP and the phenotype. Many software packages have been developed for GWAS analysis based on varied statistical models. When dealing with quantitative phenotypes, linear regression approaches are usually applied, often based on Generalised Linear Models (GLMs) and Mixed Linear Models (MLMs). MLMs are also known as Linear Mixed Models (LMMs) in the literature. In this work, we will also refer to them as LMMs.

Many GWAS packages have been developed. Some well-known ones include PLINK (Purcell et al., 2007), FaST-LMM (Lippert et al., 2011), BOLT-LMM (Loh et al., 2015), TASSEL (Bradbury et al., 2007), GAPIT (Lipka et al., 2012), GenABEL (Aulchenko et al., 2007) and GCTA (Yang et al., 2011). PLINK is a tool set for GWAS and population-based linkage analyses that provides rapid computation for large biological datasets. It was one of the earliest packages for GWAS and is viewed as a standard method. FaST-LMM applies a factorised log-likelihood function within a LMM and claims to support expanded data size and increased computational speed. BOLT-LMM utilises an efficient Bayesian mixed-model and claims to increase association power and computation speed for large datasets. TASSEL can utilise both GLMs and LMMs in determining associations, and takes into account the population and family structure. GAPIT is an R package implementing a compressed LMM. It can perform both GWAS

and genomic prediction/selection. GCTA is a tool for genome-wide complex trait analysis. It fits, by a LMM, the contribution of all SNPs as random effects, and addresses the "missing heritability" problem of human genomes. GenABEL is an R library implementing effective storage and exploration of genome-wide data, with efficient procedures for genetic data quality control. Among all these applications, we find PLINK, FaST-LMM, TASSEL and GAPIT to be widely used in the literature. The latter two are particularly popular in plant genome analysis.

Different GWAS programs often produce dissimilar association results. Newer proposed approaches usually claim to have increased statistical power over previous methods. However, there are few independent evaluations of GWAS programs, especially in a plant and plant genomics context. It is important to understand the behaviour of these popular GWAS programs, and their sensitivity to input data quantity and to phenotypes where varying numbers of biological replicates are observed.

When considering the effects of input data, intuitively, better results and higher statistical power would be expected with more data (samples). However, since genotyping can be resource intensive, there is pressure to limit the sample size while still achieving satisfying results.

Power analysis of GWAS is often used to provide insight into choosing sample sizes (Klein, 2007; Spencer et al., 2009). Statistical power is the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true. In GWAS, it reflects the ability to identify genetic variants that are genuinely associated with phenotypic variations (Teo, 2008). For real-experiment data, power analysis usually relies on many factors, including knowledge of Linkage Disequilibrium (LD), which SNPs have been genotyped and selected within the LD for GWAS (called tagged SNPs), and a good representative sample. Since it is hard to control each of these factors precisely (e.g. a perfect representative sample), results from a power calculation may be difficult to obtain and may not reflect what would actually happen in real experiments. Therefore, it is worthwhile to take a practical approach to examine how input data quantity (sample size) influences GWAS results in real experiments.

In experimental designs in plant science, it is common to see biological replications of samples; for example, 3 replicates planted for each genotype. However, the phenotypes of these replicates could be different even though their genomes are the same. As well, actual numbers of replicates can vary due to missing data. For instance, it might be that 3 replicates of each of genotypes A and B are planted, but due to uncontrollable circumstances only 2 replicates of genotype B are ultimately available. Dealing with data containing varied numbers of biological replicates in phenotypes is therefore a practical issue, and it is important to understand the effect on GWAS output of different approaches to handling it.

In this study, we investigate the effects of the amount of input data on GWAS results in the context of plants. Four widely used GWAS programs, PLINK, TASSEL, FaST-LMM, and GAPIT are used. GWAS program performance comparisons exist in the literature (Galesloot et al., 2014; Eu-Ahsunthornwattana et al., 2014), but none of them are focused on input data quantity or plant genomes. To conduct this study, we use real-data sets from a well-studied model plant, *Arabidopsis thaliana*, as well as synthetic datasets. Population structures of the *Arabidopsis thaliana* datasets are inferred and served as (optional) input when applicable. Findings of this study provide guidance on GWAS program selection, on determining how much data is necessary to yield significant results for downstream analysis, and on experimental design with respect to biological replicates.

2 Method

To address the research question of how the quantity of input genome information influences GWAS output, we use the following procedure. For a given dataset D having d samples (with either real or synthetic data), we gradually reduce the input data amount by ratio r $(0<r<1)$ to produce $t-1$ subsets of D, where $r=\frac{1}{t}$. For example, when $r=0.25$ and $t=4$, the input data is reduced by 25% each time, yielding 3 subset sizes, 25%, 50%, and 75% of the original. For each subset of size $r_i d$ $\left(r_i=\frac{i}{t}, i=1,2,...,t-1\right)$, k sets are generated by random sampling without replacement from D. Therefore, a total of $k(t-1)$ subsets of D are generated. When $r_i d$ is not an integer, its ceiling $(\lceil r_i d \rceil)$ is used. 100% of the original set is D itself. All these sets ($k(t-1)$ randomly generated subsets and D itself) are then used to perform GWAS. Prior to the actual analysis, the genotype data is filtered, removing any SNPs with minor allele frequency < 0.05. Population structure determined by each subset, denoted as K_{set}, is inferred and served as (optional) input when applicable.

In a GWAS experiment, SNPs along with their *p*-values are typically output. *P*-values are used to evaluate the significance of SNPs associated with phenotypes of interest. Since tens to hundreds of thousands of SNPs are tested simultaneously in GWAS, there arises the multiple testing problem (Noble, 2009; Pollard et al., 2005). To address the problem Bonferroni correction is sometimes used. However, the Bonferroni method is generally considered to over-correct results and be too conservative. Therefore, we apply an alternate approach, *q*-value correction (Storey and Tibshirani, 2003), which is a False Discovery Rate (FDR) based adjustment for *p*-values.

From GWAS output, SNPs with significant *p*-values (or *q*-values) are used for downstream analysis. When selecting these SNPs, either a predefined threshold (such as $p < 10^{-5}$ or $q < 0.05$) is set and all SNPs with a *p*- or *q*-value lower than that threshold are selected, or a predefined number, e.g. 20, is set and that number of SNPs with lowest *p*-values (or *q*-values) are selected for further analysis.

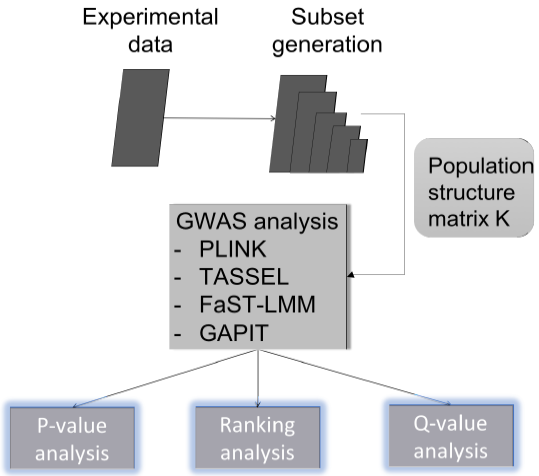
To compare the GWAS output of different runs, we consider several measures based on the following information for each reported SNP: *p*-value, *q*-value, and SNP ranking. The first two measures are the numbers of SNPs passing given thresholds p and q for *p*- and *q*-values. They are denoted as n_p and n_q, respectively. The *q*-value is calculated from the *p*-value based on the method of Storey and Tibshirani (2003) and is implemented in R. Next, the 20 most significant SNPs (SNPs with the lowest *p*-values for each run) are compared. Their *p*-values are transformed by a $-log_{10}$ function and visualised in box plots to illustrate the effects of reducing input quantity on *p*-values. Finally, the *p*-values for all SNPs are used to determine the similarity in ranking of SNPs between the output lists from a subset and the whole data D. Kendall rank correlation coefficient (denoted as τ) is used for this last measure. Its calculation is implemented in Python 2.7.

Kendall coefficient τ evaluates the degree of similarity between two ranked lists of the same objects. τ is between 0 and 1. Larger τ indicates a higher similarity of two

rankings, and $\tau = 1$ means the exact same ranking. Here, the objects are SNPs, and the ranking criterion is their *p*-values. Pearson's correlation coefficient is not applicable in this case because it measures the strength of linear correlation between two sets of data and is sensitive to outliers. There is no evidence that the *p*-values from orderings of the same SNPs would be linearly correlated, and it is possible that some of the *p*-values are outliers. Therefore, the Kendall's rank correlation coefficient is used.

In our experiment, k random sets are generated for each subset size $r_i D$, and each produces its own n_p, n_q and τ. The mean and standard deviation of the k values is calculated to represent the result for $r_i D$. The overall workflow is summarised in Figure 1.

Figure 1 Workflow of the study



3 Experiments

In this study, two kinds of data are used. One is a synthetic dataset generated using PLINK (Purcell et al., 2), and the other is a group of *Arabidopsis thaliana* datasets with multiple phenotypes. PLINK is excluded from the synthetic data comparison to avoid potential bias. All datasets have quantitative phenotypes, and their summary is listed in Table 1.

Table 1 Summary of datasets used in the experiments

Dataset name	*Number of SNPs*	*Number of samples*	*Selected phenotype*	*Biological replicates?*
AtOil	214,051	1100	18_2	Yes, up to 3
AtPolyDB	214,051	195	FT10	No
Simu_data	200,100	2000	Simulated	No

3.1 Synthetic data

The synthetic dataset, named Simu_data, is generated by PLINK using its built-in function "--simulate-qt" for quantitative traits. It contains 2000 samples across 200,100 SNPs. Among all SNPs, 100 are causal. All simulated SNPs (200,100) are unlinked. The lower and upper allele frequencies are set to 0.05 and 0.95, respectively. The generated dataset is in PLINK file format (.ped and .map), and transformed to HapMap format using TASSEL.

Other synthetic data generators such as AlphaDrop (Hickey and Gorjanc, 2012) and GPOPSIM (Zhang et al., 2015) can be used. However, our experience is that these alternatives are either difficult to use due to a lack of clear documentation regarding various parameters, or because they have portability issues. In contrast, the PLINK simulation function is easy to use and the implementation is portable and stable.

3.2 Arabidopsis thaliana data

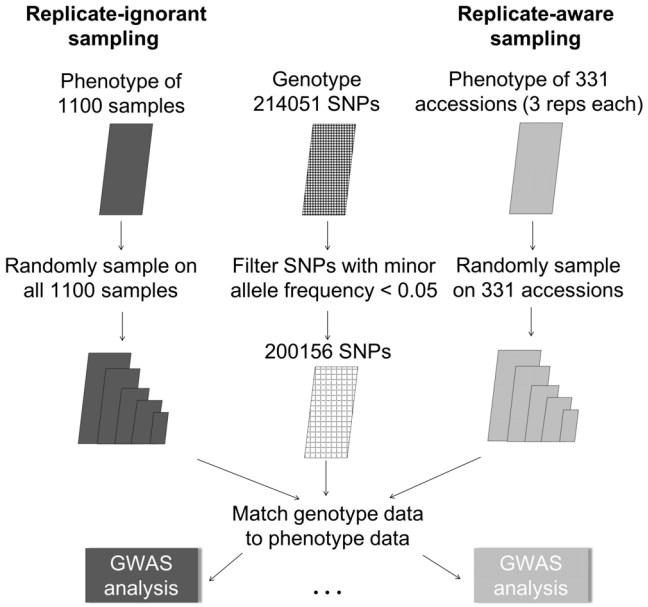
Two *Arabidopsis thaliana* datasets, named AtOil and AtPolyDB, are used. AtOil is obtained from a study of the seed oil composition of *Arabidopsis thaliana* (Branham et al., 2015). The entire dataset has 391 records (accessions) from the study. The genotype data contains 214,051 SNPs for each accession, and the phenotype data contains the relative proportions of nine principle fatty acids in *Arabidopsis thaliana* seed oil and four composite traits related to oil quality. The accessions have up to three biological replicates each, for a total of 1100 samples. The phenotype 18_2, which represents the total proportion of polyunsaturated fatty acids, is used in this work. This phenotype is chosen from among all nine fatty acid phenotypes because it has the largest variation across samples.

The AtPolyDB dataset is available from the easyGWAS website (http://easygwas.ethz.ch), and originated from two papers describing GWAS on *Arabidopsis thaliana* data (Atwell et al., 2010; Horton et al., 2012). It has 1307 samples with 214,051 SNPs each and 107 different phenotypes, though not every sample has 107 phenotypes. Phenotype FT10 (quantitative) is selected for this study since it has the largest sample size (195).

3.3 Sampling strategy

For all three source datasets, random sampling without replacement on all samples is used to generate subsets. For the AtOil dataset, an additional sampling strategy is applied. The naive random sampling on all 1100 samples might result in bias since some of the selected samples might be replicates. Therefore, a second strategy of random sampling from only the 331 accessions having three biological replicates is used. In this case, when an accession is selected, the phenotype values of all three replicates are selected, and the number of samples is always a multiple of three. The resulting subsets are balanced. The first strategy is denoted as replicate-ignorant sampling and the second as replicate-aware sampling. A diagram of performing GWAS using the two sampling strategies on AtOil is shown in Figure 2.

Figure 2 GWAS workflow on the AtOil data with two sampling strategies



3.4 Parameters

Parameters used in experiments are summarised in Table 2. The *p*-value threshold of 10^{-5} is an often used cut-off value in the literature. GWAS programs settings are listed in Table 3. The population structure matrix of each experimental dataset is interpreted using the STRUCTURE program (Pritchard et al., 2000) with default parameters and settings.

Table 2 Values and meanings of the parameters used in the experiments

Parameter	*Explanation of parameter*	*Value*
t	number of different subset sizes	10
r	reduced input ratio	0.1
k	number of random generatedsubsets per size	30
p	*p*-value threshold	10^{-5}
q	*q*-value threshold	0.05

Table 3 GWAS program settings in the study

	PLINK	*TASSEL*	*FaST-LMM*	*GAPIT*
GWAS model	Quantitative trait association	LMM (mlm module)	LMM (single-snp)	LMM
Population structure used?	No	Yes	Yes	Yes
Kinship used?	No	Yes	No	Yes

4 Results and discussion

In this section, we report the experimental results for all three datasets and compare the performance differences between programs in detail. The report starts with the AtOil dataset and two sampling strategies followed by a comparison between two *Arabidopsis thaliana* datasets, and finally the synthetic datasets.

4.1 Comparison of sampling strategies

In this subsection, the AtOil dataset is used to investigate the sampling strategy's impact on GWAS results. Original *p*-values output from each program are used as the results. *Q*-values are studied in the next subsection. First, numbers of SNPs with *p*-value $< 10^{-5}$ using replicate-aware sampling and replicate-ignorant sampling are shown in Figures 3 and 4, respectively. These figures show that sampling strategies impact the results significantly. For PLINK, the replicate-aware strategy produces many more SNPs than the replicate-ignorant one when using the same amount of input. TASSEL demonstrates a similar trend. FaST-LMM shows the opposite effects; the replicate-ignorant case outputs more SNPs. GAPIT, different than all the others, shows little variation between the two strategies.

One noticeable difference among the four GWAS outputs is that PLINK generates many more SNPs with *p*-value $< 10^{-5}$ than the other three with the same input. Further examination of the results reveals that *p*-values of the same SNPs from PLINK are much lower than from the other programs by orders of magnitude; for example, 10^{-30} versus 10^{-8}. This explains why PLINK outputs a vastly larger number of SNPs given the same *p*-value cut-off. The statistics model difference between PLINK (linear regression with Wald statistic) and the rest (LMM) is the likely key contributor to the dramatic difference in *p*-values.

Figure 3 Average numbers of SNPs with *p*-value $< 10^{-5}$ for AtOil data using the replicate-aware sampling strategy. Secondary *y*-axis (in red) is for PLINK results

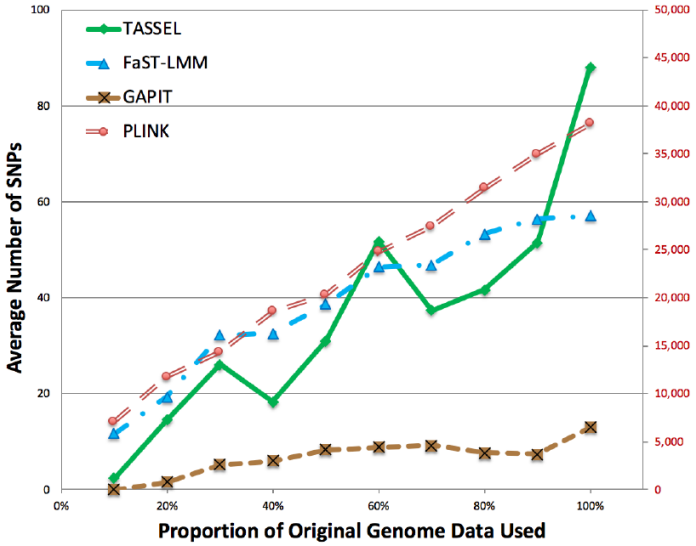
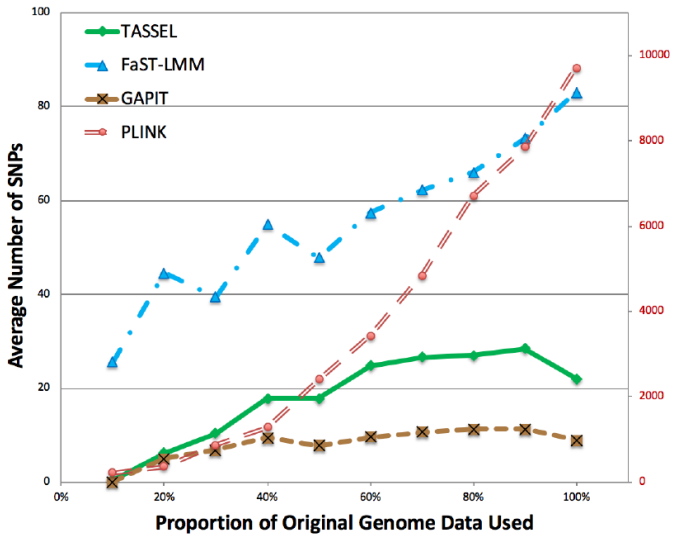


Figure 4 Average numbers of SNPs with p-value < 10^{-5} for AtOil data using the replicate-ignorant sampling strategy. Secondary y-axis (in red) is for PLINK results



TASSEL shows a decrease in the number of reported SNPs going from 90 to 100% of the data in Figure 4. However, the decrease may be an artefact. The result for the 90% subset is the average across 30 randomly generated subsets, while that for 100% is for a single dataset. More detailed results for TASSEL, specifically means and standard deviations for different subsets, are given in Table 4. (We do not show error bars in Figure 4 to prevent overcrowding.) As seen from the table, the drop from 90 to 100% of the data might be due to the variance among the 90% replicates rather than suggesting that more data leads to fewer SNPs.

Table 4 Means and standard deviations of the numbers of SNPs with p-value < 10^{-5} from TASSEL for AtOil data using replicate-ignorant sampling. "NA" indicates that there are no SNPs output for this subset size

Percentage of data	*Mean*	*Standard deviation*
100%	22.00	0.00
90%	28.45	6.82
80%	27.05	8.88
70%	26.60	8.79
60%	24.75	6.45
50%	17.85	5.52
40%	17.90	8.80
30%	10.35	4.89
20%	6.25	2.80
10%	NA	NA

We next investigate the effect of sample size on the *p*-values of the most significant SNPs reported by each program. The negative $\log_{10}$ of the *p*-values of the 20 most significant SNPs of each GWAS run are plotted in box plots. Figures 5 to 8 demonstrate the results using the replicate-aware sampling strategy for PLINK, TASSEL, FaST-LMM, and GAPIT, respectively. Results using the replicate-ignorant sampling strategy are shown in Figures 9 to 12. For each subset size there are 30 randomly generated subsets, and for each subset, the 20 most significant SNPs are considered. Therefore, *p*-values for a total of 600 SNPs are represented in each of 9 columns of the box plot. For 100% of the data on the other hand, there is only one set of the 20 most significant SNPs plotted. Hence, there are far fewer points (1/30) shown in the column for 100% of data than in columns for the subsets.

Figure 5 Negative $\log_{10}$ of *p*-values of the top 20 SNPs from PLINK for AtOil data using the replicate-aware sampling strategy. "x" in each box represents the mean value

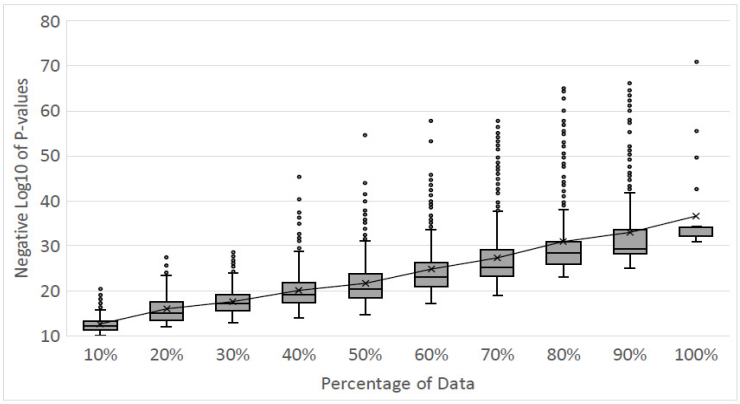


Figure 6 Negative $\log_{10}$ of *p*-values of the top 20 SNPs from TASSEL for AtOil data using the replicate-aware sampling strategy. "x" in each box represents the mean value

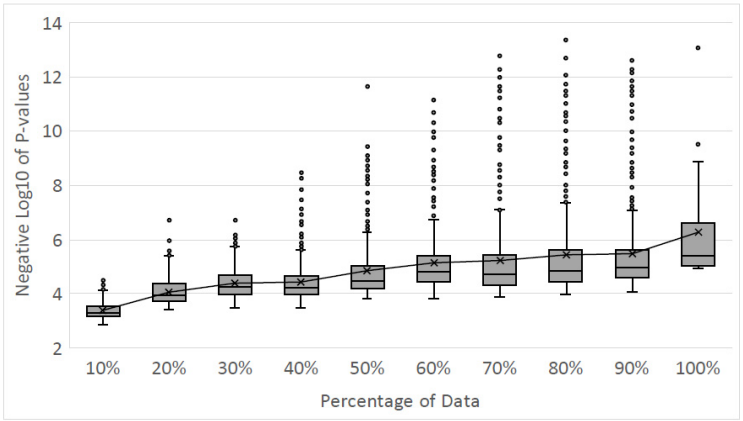


Figure 7 Negative $\log_{10}$ of *p*-values of the top 20 SNPs from FaST-LMM for AtOil data using the replicate-aware sampling strategy. "x" in each box represents the mean value

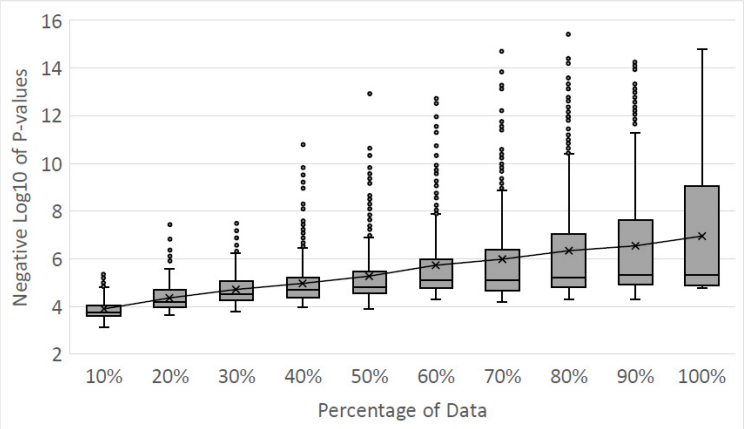
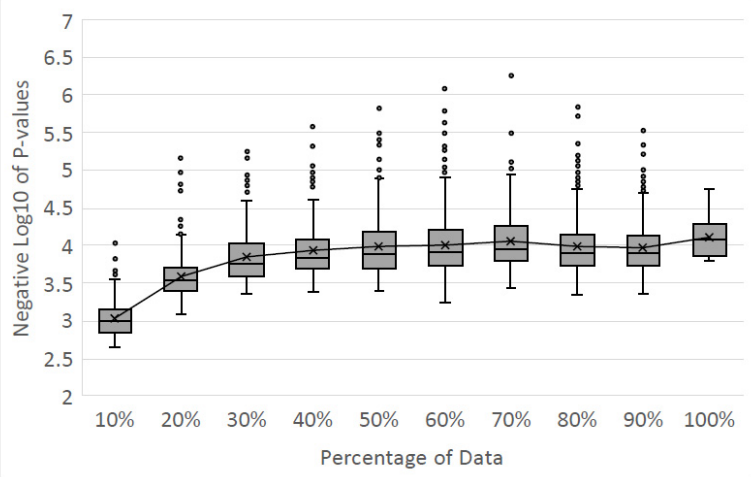


Figure 8 Negative $\log_{10}$ of *p*-values of the top 20 SNPs from GAPIT for AtOil data using the replicate-aware sampling strategy. "x" in each box represents the mean value



We first focus on the results from replicate-aware sampling. PLINK demonstrates a clear linear increase in the values of the negative $\log_{10}$ of *p*-values, which indicates that the actual *p*-values are steadily decreasing with more input data. The trend is consistent with the result shown in Figure 3 since higher negative $\log_{10}$ of *p*-values correspond to more *p*-values being less than a set threshold. FaST-LMM in Figure 7 shows a clear linear increase with greater input size as well, though variation also increases with more input. The plot for TASSEL in Figure 6 also shows a generally increasing trend for values of the negative $\log_{10}$ of *p*-values. This is consistent with the generally upward trend of the plot for TASSEL in Figure 3.

As in Figure 3, the result for GAPIT in Figure 8 is different from the results for the other GWAS programs in Figures 5 to 7. The plot in Figure 8 shows limited variation of

p-values after 30% of the input. This is consistent with the trend in Figure 3 for GAPIT where the numbers of SNPs below the *p*-value threshold stay at a relatively constant level above a subset size of 30–40%.

The box plots in Figures 9 to 12 for replicate-ignorant sampling show similar trends as for replicate-aware sampling (Figures 5 to 8), though the $-\log_{10}$ of *p*-values are generally lower. Again the plots for PLINK, TASSEL, and FaST-LMM steadily increase with greater input size, corresponding to increasing trends for these programs in Figure 4. GAPIT's behaviour in Figure 12 is again similar to that in Figure 8. PLINK shows unusual behaviour for 10% of the data in Figure 9. There is no corresponding observable behaviour for the PLINK curve in Figure 4, but this could be due to the scale of the plot; the effect is likely too small to be observed in Figure 4.

Figure 9 Negative $\log_{10}$ of *p*-values of the top 20 SNPs from PLINK for AtOil data using the replicate-ignorant sampling strategy. "x" in each box represents the mean value

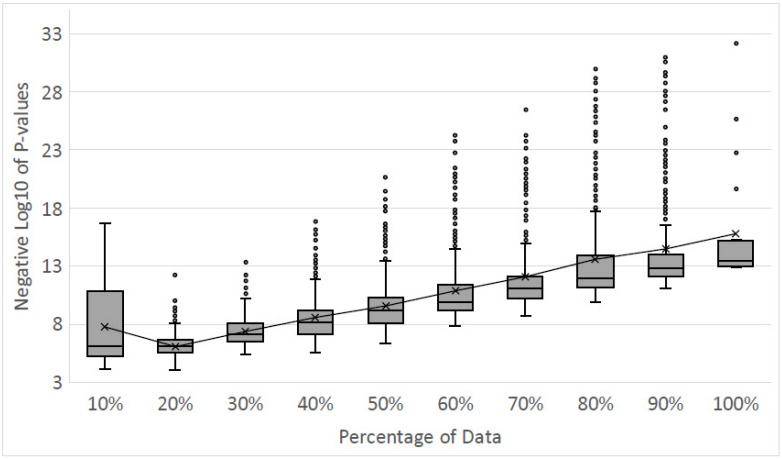


Figure 10 Negative $\log_{10}$ of *p*-values of the top 20 SNPs from TASSEL for AtOil data using the replicate-ignorant sampling strategy. "x" in each box represents the mean value.

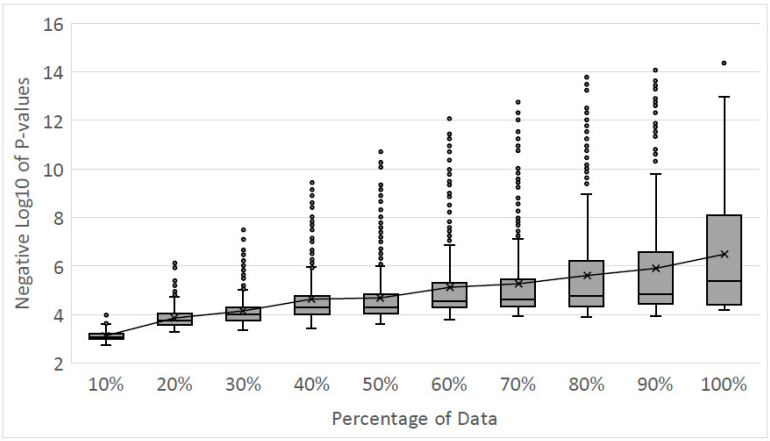


Figure 11 Negative $\log_{10}$ of *p*-values of the top 20 SNPs from FaST-LMM for AtOil data using the replicate-ignorant sampling strategy. "x" in each box represents the mean value

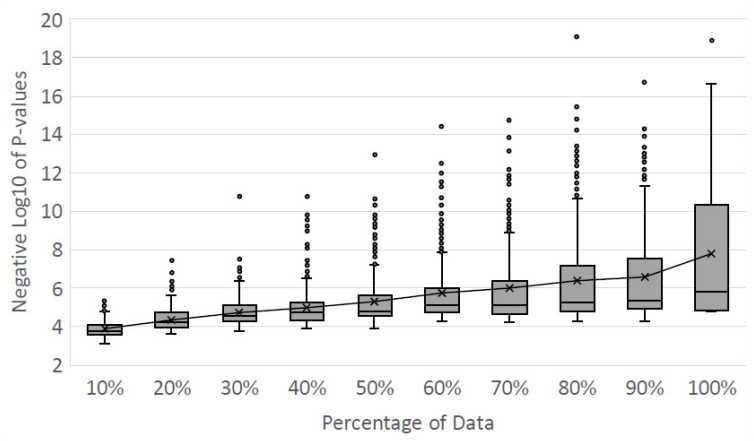
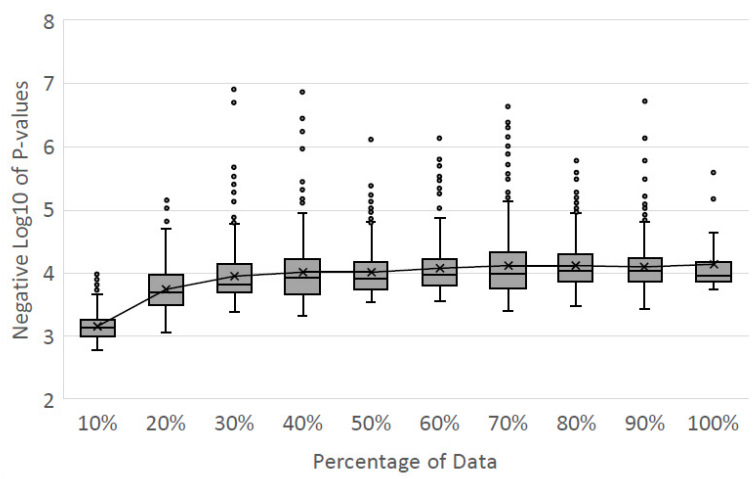


Figure 12 Negative $\log_{10}$ of *p*-values of the top 20 SNPs from GAPIT for AtOil data using the replicate-ignorant sampling strategy. "x" in each box represents the mean value



Finally, we compare the Kendall coefficient τ for the four GWAS programs using replicate-aware and replicate-ignorant sampling. Results are shown in Figures 13 and 14, respectively. The figures reveal that the choice of sampling strategy has little impact on τ. Results are almost identical except at 10% data for PLINK, and 60–70% data for TASSEL. This indicates the overall ranking of SNPs varies little between sampling strategies.

Figure 13 Average Kendall coefficient τ for AtOil data using the replicate-aware sampling strategy

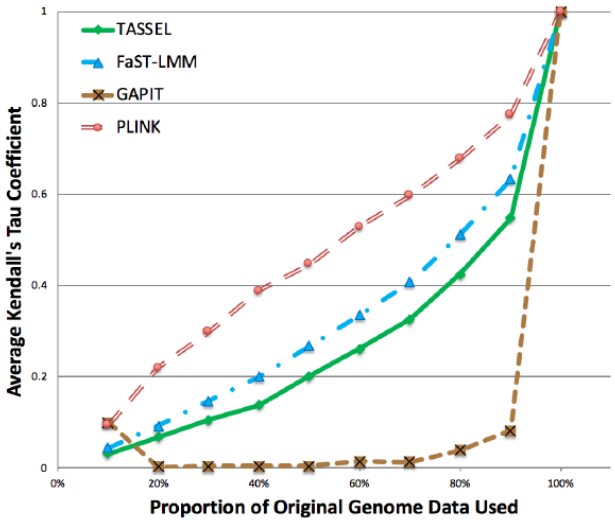
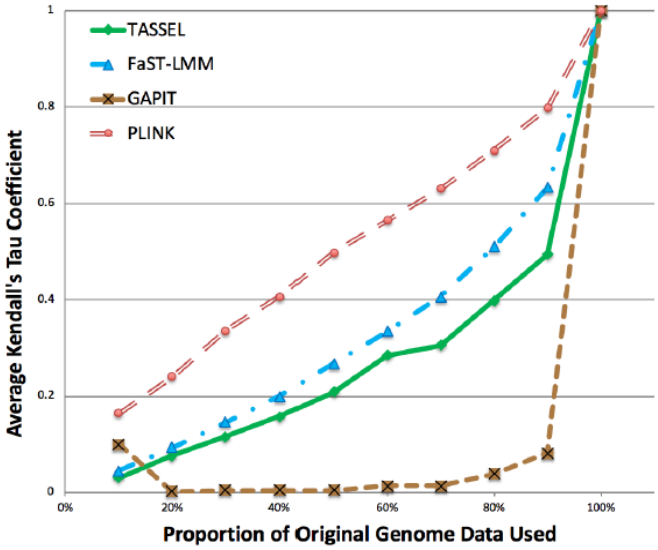


Figure 14 Average Kendall coefficient τ for AtOil data using the replicate-ignorant sampling strategy



PLINK, TASSEL, and FaST-LMM demonstrate linear-like increases in τ with more input data. GAPIT's τ values are significantly smaller than other programs' using the same subsets. Even with 90% of the original data, GAPIT's τ is below 0.1. This suggests that GAPIT is very sensitive to input genomes on this dataset. A little change of input leads to a substantial difference of output SNP lists. Further study is being carried out to investigate the reasons for this singular behaviour of GAPIT.

To conclude the above analysis, when the experimental data have varied biological replicates (one to three for this dataset), sampling strategy can heavily influence the number of SNPs passing the predefined *p*-value threshold for a given GWAS program. The replicate-aware strategy gives a balanced dataset for GWAS while the replicate-ignorant sampling strategy can give an unbalanced one. These two sampling strategies can therefore expose the effect of bias in a dataset. When the SNP selection for downstream analysis is based on predefined *p*-value thresholds, using balanced data with the same number of replicates per sample (replicate-aware strategy) versus unbalanced data (replicate-ignorant strategy) can lead to differing results, as shown in Figures 3 to 12. Of the four tested GWAS programs, PLINK, TASSEL, and FaST-LMM are more sensitive than GAPIT to the choice of balanced versus biased data.

As shown by comparing the results for PLINK versus the three other programs, the choice of *p*-value threshold for selecting significant SNPs should be program specific. A permutation test can be used to set an appropriate *p*-value threshold. It is a straightforward approach to empirically generate *p*-values related to a given null hypothesis distribution but can be time-consuming and computationally expensive. By contrast, in practice it is more common to adjust *p*-values for multiple testing using techniques such as Bonferroni or FDR-based correction. The latter will be discussed in detail in the next subsection by applying *q*-values as the correction method. It should be noted that the effects of different threshold values for significant SNPs was not examined in this study; as shown in Table 4, a constant *p*-value threshold was used.

Finally, for a given GWAS program, no clear difference of τ was observed between the two sampling strategies in our experiments. That is to say, when the selection for downstream analysis is based on rankings of SNPs, balanced versus biased data has no appreciable effect.

4.2 P-values and Q-value correction

We now investigate the effects on GWAS output of varying amounts of input based on *q*-values of reported SNPs. The results of using replicate-aware and replicate-ignorant sampling on the AtOil dataset are summarised in Figures 15 and 16, respectively. One noticeable phenomenon after the *q*-value correction is that none of the *q*-values from GAPIT pass the 0.05 threshold. We are investigating GAPIT's singular behaviour, the reasons for it, and potential methods to compensate, and will report the findings in a separate study. In this paper, we provide results from the other tested GWAS programs using *q*-value correction.

PLINK shows a linear trend with more input data, and outputs many more SNPs with *q*-value passing the threshold than other programs. Therefore, after the *p*-value adjustment, the dramatic difference seen in Figures 3 and 4 does not disappear, but is even more pronounced. This suggests that our previous recommendation to choose *p*-value thresholds in a program specific manner is also applicable to when *q*-values are used. Comparing PLINK's results between replicate-aware and replicate-ignorant sampling, the latter produced many more SNPs passing the *q*-value threshold than the former. This phenomenon is also observed in the *p*-value cases (in Figures 3 and 4).

Results for TASSEL and FaST-LMM show a generally increasing trend in Figures 15 and 16. However, in Figure 15, there is a decrease in the average numbers of SNPs with *q*-value < 0.05 for FaST-LMM when the amount of input data increases from 30 to 40%. Referring to Table 5, one can see that the standard deviation for 30% of all data is

substantially larger than the ones for other proportions. On further investigation, we find that this is due to two runs (out of the total, 30) yielding extremely large numbers of SNPs passing the threshold. The runs not only cause large standard deviation, but also inflate the average number of SNPs for 30% of the data. Therefore, the decrease (for 30 to 40% of the data) is because of a small number of outliers rather than suggesting that more data leads to fewer SNPs. The decrease seen in TASSEL for 60% of the data is attributable to the same phenomenon and explained in the same way.

Figure 15 Average numbers of SNPs with q-value < 0.05 for AtOil data using the replicate-aware sampling strategy. Secondary y-axis (in red) is for PLINK results

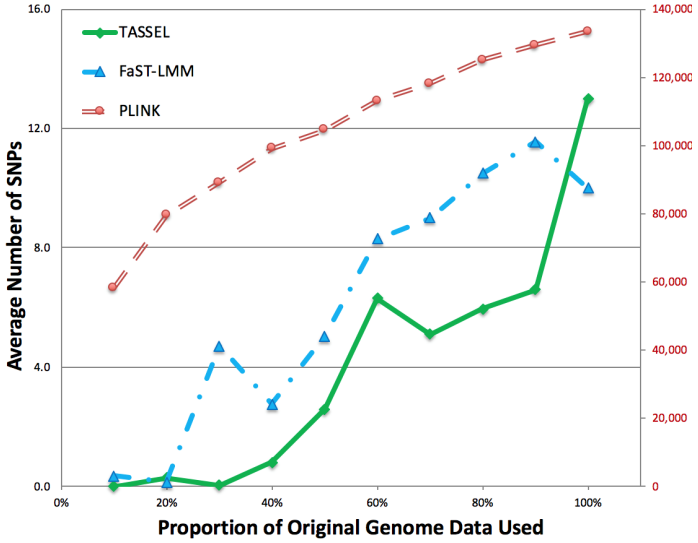


Figure 16 Average numbers of SNPs with q-value < 0.05 for AtOil data using the replicate-ignorant sampling strategy. Secondary y-axis (in red) is for PLINK results

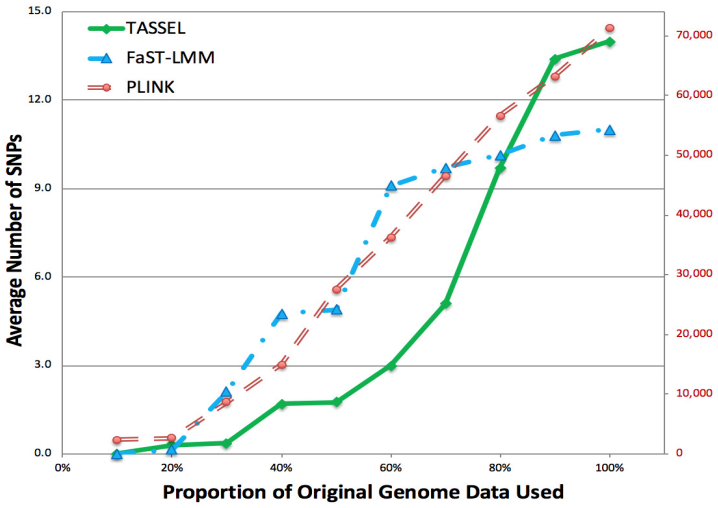


Table 5 Standard deviations of the numbers of SNPs with q-value < 0.05 for each subset size from GWAS programs using AtOil data and replicate-aware sampling. "NA" indicates that there are no output SNPs at this subset size and hence no standard deviation can be calculated

Percentage of data	*TASSEL*	*FaST-LMM*	*PLINK*
90%	5.75	4.19	2282.15
80%	3.71	3.40	3540.04
70%	3.99	4.95	5421.08
60%	10.59	4.82	6013.97
50%	2.06	2.16	9754.23
40%	1.32	2.57	8693.76
30%	0.22	12.33	8013.20
20%	1.13	0.37	9293.72
10%	NA	1.57	7601.06

It may appear that standard deviations for PLINK in Table 5 are quite large (in a scale of 10^4). However, compared to the mean values (in a scale of 10^5 to 10^6), they are still relatively small (typically within 10% of the mean).

Finally, because of GAPIT's singular behaviour upon q-value correction and because the latter is just one method for p-value adjustment, we use unadjusted p-values from all programs for the subsequent analyses. Consequently, GAPIT can again be included in the analysis. This decision to use unadjusted p-values is supported by the observation that rankings of SNPs stay the same before or after the correction, so it does not change the results for Kendall coefficient τ.

4.3 Comparison between Arabidopsis thaliana datasets

In this subsection, we compare results between two *Arabidopsis thaliana* datasets, AtOil and AtPolyDB. *P*-values output from GWAS programs and SNP rankings are used. Numbers of SNPs with p-value < 10^{-5} and Kendall coefficient τ for the AtPolyDB dataset with FT10 phenotype are shown in Figures 17 and 18, respectively. Since the plots of the negative $\log_{10}$ of p-values tend to only confirm the counts of SNPs below a given p-value threshold, the former are not shown for the AtPolyDB dataset.

From Figures 3, 4, and 17, one can see that PLINK shows similar trends as for the AtOil data: PLINK consistently generates orders of magnitude more SNPs than the other programs with the same input. GAPIT, for most cases, outputs the fewest SNPs among the four programs. GAPIT also shows a decrease after the 80% data point in Figure 17, which is not observed in Figures 3 and 4. TASSEL exhibits a linear-like relationship between the number of SNPs and amount of input data in Figure 17; the behaviour is more similar to that in Figure 4 and less to what is observed in Figure 3 (where TASSEL fluctuates more).

Figure 17 Average numbers of SNPs with p-value < 10^{-5} for AtPolyDB data with FT10 phenotype. Secondary y-axis (in red) is for PLINK results

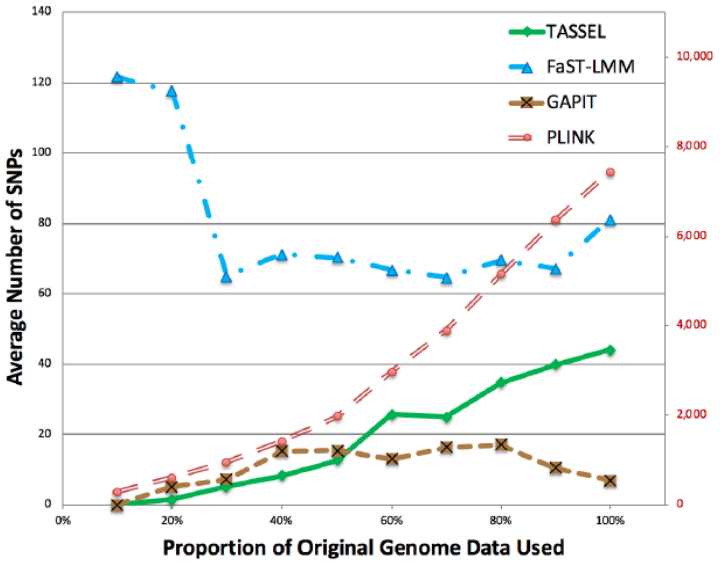
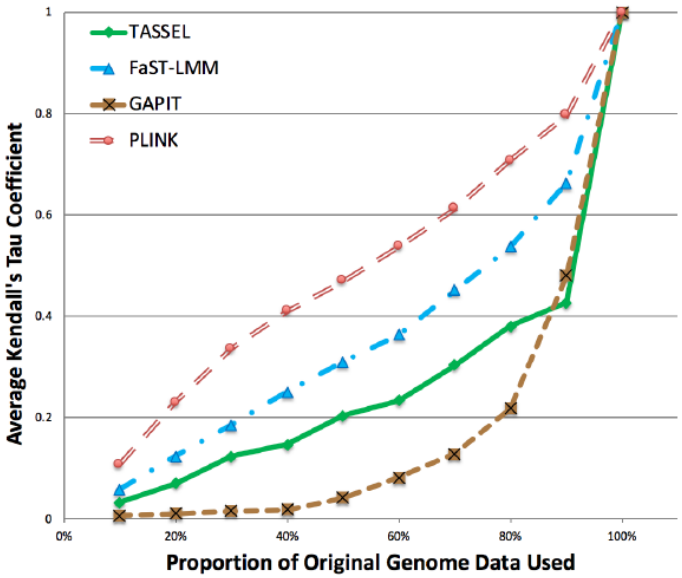


Figure 18 Average Kendall coefficient τ for AtPolyDB data with FT10 phenotype



FaST-LMM displays an unusual trend in Figure 17 compared with Figures 3 and 4. Figures 3 and 4 show generally increasing trends with more input data despite some fluctuation in Figure 4. However, in Figure 17, FaST-LMM produces vast numbers of SNPs for the 10% and 20% subsets, followed by a dramatic decrease at 30%. It fluctuates and increases marginally with more input data afterward, but never overtakes the 10%

and 20% data points, even with the whole dataset. Detailed analysis as summarised in Table 6 reveals that this behaviour is probably due to large standard deviations in the number of SNPs with p-value $< 10^{-5}$. This also implies that different sample subsets could yield varying results, and more randomly generated subsets (increasing k) could reduce the large standard deviations. One more general observation is that the AtPolyDB dataset is much smaller than the AtOil dataset; it is only about 1/6 the size of the AtOil dataset (195 versus 1100 samples). Therefore, results for the AtOil dataset are much more vulnerable to the effects of outliers caused by the random sampling.

Table 6 Means and standard deviations of the numbers of SNPs with p-value $< 10^{-5}$ from FaST-LMM for AtPolyDB

Percentage of data	*Mean*	*Standard deviation*
100%	80.00	0.00
90%	67.15	10.05
80%	69.45	11.58
70%	64.6	22.16
60%	66.85	13.63
50%	70.25	10.98
40%	71.25	18.42
30%	65	26.59
20%	117.6	48.75
10%	121.6	127.81

Figure 18 demonstrates trends similar to the AtOil cases (see Figures 13 and 14). When given the same input, PLINK produces the highest τ, followed by FaST-LMM, TASSEL, and GAPIT. That is to say, generally, PLINK has the highest SNP ranking similarity between subsets and the whole dataset. GAPIT demonstrates high sensitivity to input data amount.

To conclude, we see that the influence of input data quantity on the ranking of SNPs (as shown by τ) is consistent between the two *Arabidopsis thaliana* datasets, while for the number of SNPs below a set threshold, results are program specific. In the case of the latter criterion, PLINK demonstrates similar trends between the two datasets, and GAPIT shows the strongest sensitivity (of input quantity) among all programs. It usually generates the least number of SNPs for a given amount of input. For TASSEL and FaST-LMM, they exhibit differing results between the two datasets, especially the unusual decrease by FaST-LMM shown in Figure 17.

Results from both *Arabidopsis thaliana* datasets imply that when having small sample sizes for GWAS, PLINK is expected to produce the most similar SNP rankings as additional samples are added. Viewed in another way, PLINK can best approximate the SNP ranking for a (potentially) large population when only a small dataset is available. For other programs, especially GAPIT, rankings change dramatically with even a small change in sample size.

4.4 Comparison between synthetic and real data

In this subsection, we focus on the differences between using the synthetic dataset, Simu_data, versus the two *Arabidopsis thaliana* datasets. PLINK is excluded from Simu_data analysis to avoid potential bias since it was used to generate the data. Counts of significant SNPs and ranked SNP lists are used for comparison. Numbers of SNPs with p-value < 10^{-5} and Kendall coefficient τ of Simu_data with quantitative phenotype are shown in Figures 19 and 20, respectively.

Figure 19 Average numbers of SNPs with p-value < 10^{-5} for Simu_data with quantitative phenotype

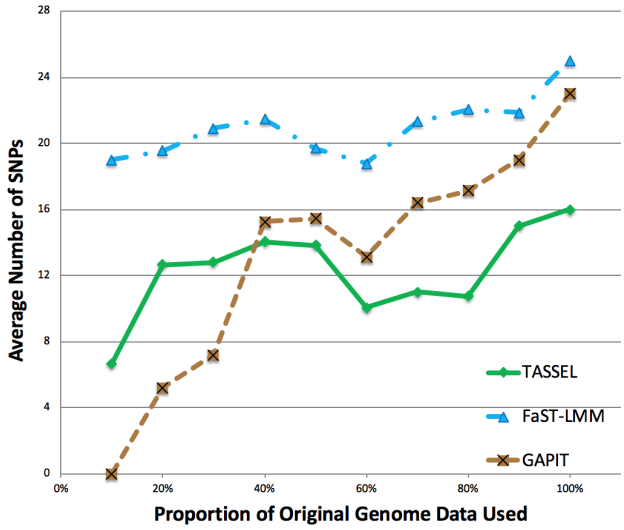


Figure 20 Average Kendall coefficient τ for Simu_data with quantitative phenotype

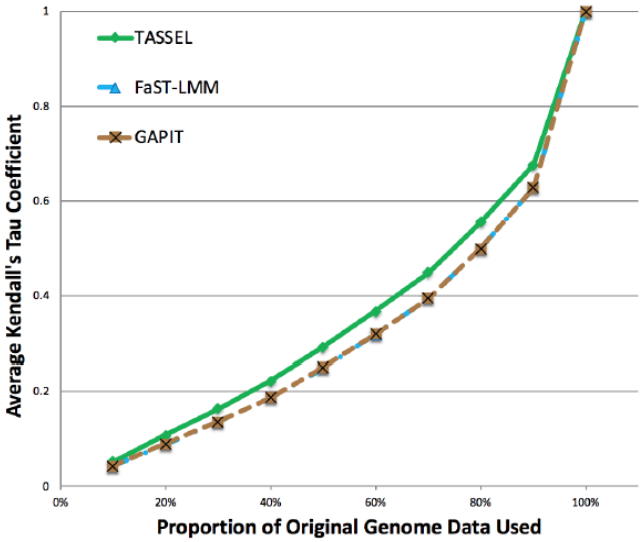


Figure 19 shows that the numbers of SNPs fluctuate for TASSEL and FaST-LMM, while GAPIT demonstrates a linear increase except from 50 to 60% of the data. Numbers of SNPs differ less for TASSEL and FaST-LMM on Simu_data (7 to 16, and 19 to 25, respectively, from 10 to 100% of total data) than on *Arabidopsis thaliana* datasets (Figures 3, 4, and 17). This suggests that for TASSEL and FaST-LMM on Simu_data, using just subsets of data could produce significant SNPs comparable to using the whole dataset.

The Kendall coefficient τ results for Simu_data (see Figure 20) show mildly exponential increases with more input, and there is only a minor difference between the programs. This is quite different from the results for the AtOil and AtPolyDB datasets where programs differ in Kendall coefficient τ. Despite the minor difference between TASSEL and the others, all three tested programs produce almost identical τ with the same input. On this dataset, use of different programs has little impact on SNP rankings.

One possible reason for this phenomenon is that some specific features of real-data sets are not accurately reflected or are hardly distinguishable in the synthetic dataset, such as population structure or kinships. SNPs in Simu_data are generated based on the assumption that they are independent, so there are no kinship relations between samples (and PLINK does not provide an option to include kinship in the data generator). Further study of this phenomenon is warranted to generate a conclusive explanation. Methods for generating synthetic data besides that provided by PLINK are also being explored.

To conclude, we see dramatic differences between the results on synthetic and real data. Many consistent observations from the real data disappear when using the synthetic data. For example, for the number of SNPs with p-value $< 10^{-5}$, GAPIT no longer demonstrates higher sensitivity to input data compared to the other programs. For the Kendall coefficient τ, differences between programs with the same input are barely discernible compared to the results in Figures 13, 14, and 18. Results from our experiments suggest that the data simulation method in PLINK might not accurately model real plant data and capture their features. Therefore, better synthetic data generators are needed in this field.

5 Conclusions and future work

This study investigated how input data quantity influences GWAS results in plant genomics. Four widely used GWAS programs, PLINK, TASSEL, GAPIT, and FaST-LMM, were compared in the experiments. Both synthetic data and real *Arabidopsis thaliana* data were used. For determining results, p-values, q-values, and SNP ranking measurements were determined. To be specific, numbers of SNPs passing given p- and q-value thresholds, and Kendall rank correlation coefficient τ between output SNP lists were used. As well, for one dataset the negative $\log_{10}$ of p-values for the 20 most significant SNPs in each run were visualised using box plots. In practice, plant samples with a varied number of biological replicates are often obtained. This situation was also explored in our experiments.

From the experimental results, we see that balanced (same number of replicates per genotype-phenotype combination) versus unbalanced (any data, no matter the number of replicates) input data affect GWAS results. These differences are most pronounced for PLINK, FaST-LMM, and TASSEL. When considering the ranking of SNPs, minimal variations between the two situations are generally observed.

PLINK consistently generates orders of magnitude more SNPs than the other programs with the same input. GAPIT, for most cases, outputs the fewest SNPs. This may be because of its alternate statistical model. GAPIT also demonstrates singular behaviour in the results of Kendall coefficient τ and even failed to report any SNPs with q-values less than 0.05. This is being investigated in detail for a conclusive explanation.

Results differ between synthetic data and real data. One possible reason for this phenomenon is that the synthetic data does not adequately reflect the characteristics of real data. Future research will explore other methods to generate synthetic data that is more similar to real data.

In terms of the effects of input data quantity on GWAS results, it is both program and measurement specific. One can expect a linearly increasing relationship between input quantity and numbers of SNPs passing a threshold for PLINK, but no guarantee for TASSEL, FaST-LMM, and GAPIT. Viewed alternatively, it appears that the latter three programs are less robust to a reduction in the amount of input. The experimental results also suggest that setting a threshold p- or q-value for significant SNPs should be program specific.

The SNP ranking generally demonstrates a linear increase in the similarity between subsets and the whole dataset with more input data. On real-data sets, PLINK achieves the highest similarity while GAPIT the lowest. Therefore, when only limited samples are available, PLINK is expected to produce a ranked SNP list that is closest to the SNP list generated for a large sample. For a program like GAPIT, rankings can change dramatically with even a small change in the sample size. Hence, researchers should conduct additional GAPIT runs whenever the amount of experimental data changes.

For situations of small sample sizes, this study suggests that PLINK is a good choice for GWAS as its performance is expected to be similar to that for a potentially larger sample size. If a GWAS method based on a LMM is desired, either of TASSEL or FaST-LMM can be used. However, GAPIT appears to be quite sensitive to input quantity and dissimilar results can be expected with even small changes in amount of input.

In addition, balanced experimental data is recommended as it is known to enhance the power of the statistical tests. When only unbalanced data is available, proper pre-processing to make it balanced is encouraged. If one has to use unbalanced data for GWAS, SNP ranking is recommended as a selection criterion over a pre-defined p-value threshold since the ranking is not heavily influenced by the unbalanced versus balanced data.

In some applications, generating and collecting genotype and phenotype data is resource intensive. There is a trade-off between acquiring more data and getting better results. This study suggests that some programs such as TASSEL and FaST-LMM could produce improved results with added data for most cases, while other programs might not. There might even be worse results with more input data quantity in some situations. Choices should be made according to the program performing GWAS analysis and the criteria for selecting SNPs for downstream study. We believe that the results from this study can provide guidance on selecting GWAS programs given varied amounts of experimental data and on understanding how the quantity of input data effects results.

References

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M. and Hu, T.T. et al. (2010) 'Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines', *Nature*, Vol. 465, No. 7298, pp.627–631.

Aulchenko, Y.S., Ripke, S., Isaacs, A. and Van Duijn, C.M. (2007) 'GenABEL: an R library for genome-wide association analysis', *Bioinformatics*, Vol. 23, No. 10, pp.1294–1296.

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*, Vol. 23, No. 19, pp.2633–2635.

Branham, S.E., Wright, S.J., Reba, A. and Linder, C.R. (2015) 'Genome-wide association study of *Arabidopsis thaliana* identifies determinants of natural variation in seed oil composition', *Journal of Heredity*, Vol. 107, No. 3, pp.248–256.

Bush, W.S. and Moore, J.H. (2012) 'Genome-wide association studies', *PLoS Computational Biology*, Vol. 8, No. 12, p.e1002822.

Eu-Ahsunthornwattana, J., Miller, E.N., Fakiola, M., Jeronimo, S.M., Blackwell, J.M. and Cordell, H.J., (2014) 'Comparison of methods to account for relatedness in genome-wide association studies with family-based data', *PLoS Genetics*, Vol. 10, No. 7, p.e1004445.

Galesloot, T.E., Van Steen, K., Kiemeney, L.A., Janss, L.L. and Vermeulen, S.H. (2014) 'A comparison of multivariate genome-wide association methods', *PLoS One*, Vol. 9, No. 4, p.e95923.

Hickey, J.M. and Gorjanc, G. (2012) 'Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods', *G3: Genes, Genomes, Genetics*, Vol. 2, No. 4, pp.425–427.

Hirschhorn, J.N. and Daly, M.J. (2005) 'Genome-wide association studies for common diseases and complex traits', *Nature Reviews Genetics*, Vol. 6, No. 2, pp.95–108.

Horton, M.W., Hancock, A.M., Huang, Y.S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N.W., Platt, A., Sperone, F.G. and Vilhjálmsson, B.J. et al. (2012) 'Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the regmap panel', *Nature Genetics*, Vol. 44, No. 2, p.212.

Klein, R.J. (2007) 'Power analysis for genome-wide association studies', *BMC Genetics*, Vol. 8, No. 1, p.58.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S. and Zhang, Z. (2012) 'GAPIT: genome association and prediction integrated tool', *Bioinformatics*, Vol. 28, No. 18, pp.2397–2399.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I. and Heckerman, D. (2011) 'Fast linear mixed models for genome-wide association studies', *Nature Methods*, Vol. 8, No. 10, pp.833–835.

Loh, P-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M. and Berger, B. et al. (2015) 'Efficient Bayesian mixed-model analysis increases association power in large cohorts', *Nature Genetics*, Vol. 47, No. 3, pp.284–290.

Noble, W.S. (2009) 'How does multiple testing correction work?', *Nature Biotechnology*, Vol. 27, No. 12, pp.1135–1137.

Pollard, K.S., Dudoit, and S. van der Laan, M.J. (2005) 'Multiple testing procedures: the multtest package and applications to genomics', *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, pp.249–271.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) 'Inference of population structure using multilocus genotype data', *Genetics*, Vol. 155, No. 2, pp.945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I. and Daly, M.J. et al. (2007) 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *The American Journal of Human Genetics*, Vol. 81, pp.559–575.

Spencer, C.C., Su, Z., Donnelly, P. and Marchini, J. (2009) 'Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip', *PLoS Genetics*, Vol. 5, No. 5, p.e1000477.

Storey, J.D. Tibshirani, R. (2003) 'Statistical significance for genomewide studies', *Proceedings of the National Academy of Sciences*, Vol. 100, No. 16, pp.9440–9445.

Teo, Y.Y. (2008) 'Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure', *Current Opinion in Lipidology*, Vol. 19, No. 2, pp.133–143.

Wang, M., Yan, J., Zhao, J., Song, W., Zhang, X., Xiao, Y. and Zheng, Y. (2012) 'Genome-wide association study (GWAS) of resistance to head smut in maize', *Plant Science*, Vol. 196, pp.125–131.

Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) 'GCTA: a tool for genome-wide complex trait analysis', *The American Journal of Human Genetics*, Vol. 88, No. 1, pp.76–82.

Zhang, Z., Li, X., Ding, X., Li, J. and Zhang, Q. (2015) 'GPOPSIM: a simulation tool for whole-genome genetic data', *BMC Genetics*, Vol. 16, No. 1, p.10.