EFFICIENT GENOTYPING BY SAMPLING EXTREME INDIVIDUALS IN A GENOME WIDE ASSOCIATION STUDY IN PLANTS

by

Wenqian Kong

(Under the Direction of Paul Schliekelman)

Abstract

We evaluated statistical power of selective genotyping strategies based on sampling extreme individuals a genome-wide association study (GWAS). Simulation with a theoretical set up and with application in the actual data-set provides guidance on determining the minimum individuals from the extremes needed to detect causal variants to reach 80% statistical power. We compared power and false discovery rates of three different methods in a real-world sorghum diversity panel using Fisher's exact test, analysis of variance (ANOVA) and a popular software GAPIT which applies mixed model for variant detection and controls for population structure. Our simulation results also discover that the power of detecting causal SNP markers in selective genotyping is dependent on the initial population size. This strategy is particularly helpful in genetic studies to reduce genotyping costs for variant detection and validation.

Index words: GWAS, sampling, bulk-segregant analysis (BSA), power analysis

EFFICIENT GENOTYPING BY SAMPLING EXTREME INDIVIDUALS IN A GENOME WIDE ASSOCIATION STUDY IN PLANTS

by

Wenqian Kong

B.S., Yangzhou University, 2010

M.S., University of Georgia, 2013

Ph.D.,University of Georgia, 2017

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Master of Science

Athens, Georgia

2018

©2018

Wenqian Kong

All Rights Reserved

EFFICIENT GENOTYPING BY SAMPLING EXTREME INDIVIDUALS IN A GENOME WIDE ASSOCIATION STUDY IN PLANTS

by

Wenqian Kong

Major Professors:	Paul Schliekelman
Committee:	Jaxk Reeves
	Cheolwoo Park

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2018

EFFICIENT GENOTYPING BY SAMPLING EXTREME INDIVIDUALS IN A GENOME WIDE ASSOCIATION STUDY IN PLANTS

Wenqian Kong

April 22, 2018

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Introduction and Literature Review

Selective genotyping is a sampling design for genetic studies. When limited numbers of individuals can be genotyped (for example, due to cost constraints), appropriate sampling methods presumably provide adequate statistical power to detect genetic variants associated with traits of interest[1, 2]. In contrast to random sampling, selecting individuals at extreme tails of a phenotypic distribution was first proposed for genetic mapping in a quantitative trait loci (QTL) study pioneered by Lander and Botstein [3], due to the fact that individuals with extreme phenotypes provide more linkage information than random individuals.

A similar idea was experimentally applied in "bulked segregant analysis" (BSA) of mapping populations. BSA involves genotyping of pools of DNA based on phenotypes, usually of individuals with extreme trait values from segregating populations and searching for genetic markers associated with the trait of interest [4]. BSA is best suited to discrete traits, but may be applicable to QTLs of particularly large phenotypic effect [5]. Recently, next generation sequencing-based BSA has been applied to detecting major QTLs in bi-parental populations for resistance of rice against fungal pathogens and seedling vigor [6], grain protein content

in wheat, and rust resistance in soybean [7].

Recently, genome wide association studies (GWAS) have emerged as a valuable complement to QTL mapping, offering much finer resolution but at a high false-positive rate and with lower power in detecting rare variants [8]. The selective genotyping method has recently been applied in association studies [9], and is especially attractive in detecting rare causal variants [10, 11, 12], mitigating constraints due to limited sample sizes and explaining some missing heritability [13]. Intuitively, sampling the extreme phenotypes of a population should enrich the frequency of rare alleles and thus increase the power of detection [14, 15], as demonstrated in the extreme discordant sib-pair design [16].

Selective genotyping has been widely applied in human genome-wide association studies (GWAS) experiments, but seldom employed in plants [9]. The design of most GWAS experiments in plants are somewhat different than those in humans: GWAS in humans usually involves hundreds of thousands of heterozygous loci while many crops have mostly homozygous loci due to self-fertilization [17, 18, 19]. Therefore, the parameters used and the sampling strategy might be different in plants.

In this thesis, we investigate optimization of number of individuals with extreme phenotypes of quantitative traits to select from plant-based GWAS experiments and estimate the statistical power under a range of scenarios such as different allele frequencies, variance explained, and multiple alleles affecting a trait. This work would benefit researchers to validate their initial GWAS result, increase power for detecting rare variants and reduce costs of genotyping.

Chapter 2

Methods

2.1 Theoretical setup

We assume that each observation consists of m single nucleotide polymorphism (SNP) markers, which follows a independent Bernoulli distribution:

$$X_j \sim Bernoulli(p_j) \text{ for } j = 1, 2, \ldots, m \tag{2.1}$$

the success probability p_j is the allele frequency of the jth marker. In this case, we consider the situation that there are only two genotypes and both are homozygous, a situation which is reflected in many GWAS experiments in plants. One response variable was generated based on a total of k explanatory variables, each explaining r_i of the total variance (R^2), allele frequency p_i, and with an additive effect equals to β_i. The other $m-k$ variables (SNPs) were considered noise, with additive effects of zeros. In more detail,

$$y_i = \beta_0 + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + ... + \beta_m \times x_{im} + \epsilon_i;\ i = 1, 2, \ldots, n, \tag{2.2}$$

where

For this $n \times m$ data-set, the typical method to detect if each marker is significantly associated with the trait is to use a two-sample t-test for each marker. The non-central parameter for this two sample t-test is:

$$\lambda = \frac{1}{\sqrt{(var(\epsilon_i) * (\frac{1}{np_i} + \frac{1}{nq_i})}} = \frac{1}{\sqrt{\frac{1-R^2}{nR^2}}} \tag{2.3}$$

Therefore, the theoretical power for the two sample t-test is:

$$power = 1 - T_{v,\lambda}(t_{1-\alpha/2}) + T_{v,\lambda}(-t_{1-\alpha/2})$$

where $t_{1-\alpha/2}$ is the upper $\alpha/2$ quantile, λ is the non-central parameter, v is the degree of freedom, and T is the non-central t-distribution CDF.

The power of detecting the marker at different alpha levels are shown in Table 2.1. As is suggested in the equation (2.3), the power calculation should not be influenced by the allele frequencies. As the R^2 increases, the power of detecting SNP makrer increases; and as α decreases, the power of detecting SNP marker decreases.

Table 2.1: Theoretical power of detecting SNP marker at different alpha levels with n=1000 and m=100 in a two-sample t-test

R^2	λ	Power				
		0.001	$\alpha = 5 \times 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-6}$	$\alpha = 10^{-7}$
0.01	3.1782	0.4553	0.3810	0.1386	0.0433	0.016
0.05	7.2548	1.0000	0.9999	0.9986	0.9909	0.9730
0.10	10.5409	1.0000	1.0000	1.0000	1.0000	1.0000

2.2 Single SNP simulation with extreme individuals

To simplify the case, we simulated a total of $m = 100$ SNP markers with one SNP marker explaining a total of 1%, 5%, and 10% of total variance, respectively, and the other 99

markers were considered as 'noise' to evaluate for false positives. Therefore, the equation (2.1) was simplified to

$$y_i = \beta_0 + \beta_1 \times x_{i1} + \epsilon_i,\ i = 1, 2, \ldots, n, \tag{2.4}$$

We simulated the variance of ϵ based on $\beta_1^2 \times p \times q \times \frac{(1-R^2)}{R^2},\ q = 1-p$ to assign the value of R^2.

Next, the data-set was subsampled by choosing the extremes of the response variable at both tails. The sizes of subsets (n_{sub}) ranged from $n_s = 10$ to $n_l = 100$ from each tail, incremented by $s = 5$ individuals. Since the selected sample fails to meet the assumptions of linear regression, we used Fisher's exact test by transforming the selected response variable into binary forms and tested the significance for each explanatory variable. The upper 50% of the response variable was assigned to 1 and lower one to 0. This simulation program was performed 1000 times. The allele frequencies were taken at $p = 0.1$ and $p = 0.5$ The empirical power was calculated as the proportion of times the explanatory variable detected significance at alpha levels of 10^{-5}, $10^{-6}$and 10^{-7}. For each simulation, we plotted the power against subsample sizes.

Increasing the significance threshold, as expected, decreases the power of detecting SNP markers. The significance threshold at 10^{-7} was calculated based on Bonferroni correction using total number of filtered markers ($0.05/100000 = 5\times10^{-7}$). However, in a GWAS experiment, nearby markers are more correlated than distant markers, usually as a result of linkage disequilibrium (LD), which is the non-random association between loci. Over time, recombination shuffles genetic materials between chromosomes resulting in LD decay. LD decays gradually and proportionally to the rate of recombination over time. In sorghum, the average LD decays on background level is within 150 kb [20], and the size of the sorghum genome is approximately 730 Mb [21], suggesting that there are approximately 4867 ($730Mb/150kb$)LD

blocks in the sorghum genome. This also suggests that using a threshold of 10^{-5} (0.05/4867) is appropriate for a sorghum or other similar GWAS experiments.

2.3 Actual data-set

The actual data-set came from a US sorghum association panel (SAP) consisting of 354 individuals. Phenotypic data was collected from 354 and 349 individuals in 2009 and 2010 respectively, but data from 2009 is used in this thesis. A total of 265,487 raw published SNP data were generated based on genotyping-by-sequencing (GBS) [19]. We deleted a total of 8,533 SNP markers with more than 50% missing data, 111,711 markers with minor allele frequency smaller than 5% and 27,589 redundant markers (similarity greater than 90%), and retained 117,654 markers for the analysis. Data filtering used GenABEL package [22]. The original genotype file had approximately 18.7% missing data. We imputed the missing data using nearest 3 neighboring SNPs based on physical distance and passed the result to filtered genotype file. Simulation used the R scrime package [23].

2.4 Multiple SNP simulation with extreme individuals

We conducted a simulation scenario using a total of 100 SNP markers with 4 SNPs influencing a trait, and the other 96 SNPs considered as "noise". This simulation was based on a previous study in sorghum suggesting four general loci for plant height, *DW1- DW4* [24], which explain a considerable amount of variance in plant height of sorghum (Table 2.2). A previous GWAS study detected a total of four peak SNPs associated with plant height with data collected in 2009 (Table 2.2). The mutiple SNP simulation used the similar parameters compared to the actual case (Table 2.3). In detail, a total of 100 SNPs were simulated with four explaining 5%, 5%, 10% and 20% of the variances and with allele frequencies (p) 0.07,

Table 2.2: Summary of four peak SNP markers in the imputed data-set from the actual GWAS experiment

SNP	Var explained	Allele freq (p)	P-values (Simple regression)	P-values (Multiple regression)
S4_52707130	7.91%	0.14	2.00e-05	0.0017
S6_42138323	13.43%	0.30	5.21e-12	3.546e-09
S7_56824744	7.47%	0.06	8.104e-06	3.475e-05
S9_57236778	22.48%	0.36	$<$ 2.2e-16	$<$ 2.2e-16

0.15, 0.30 and 0.40 respectively (Table 2.3). The sampling strategy and statistical test were the same as single SNP simulation.

Table 2.3: Parameters used for multiple SNP simulation

SNP	Var explained	Allele freq (p)	β_i
x_1	5%	0.07	0.8764
x_2	5%	0.15	0.6262
x_3	10%	0.30	0.6900
x_4	20%	0.40	0.9128

2.5 Single SNP simulation with the actual data

We selected one SNP marker S2_64198488 with allele frequency 0.469, and assumed that this SNP explained about 10% of the total variance for a trait. We simulated the trait using information from this SNP marker, and then used the same extreme sub-sampling strategy as before to evaluate the result. GWAS analysis used the compressed mixed linear models (CMLM) [25] for the total of 15,222 markers on chromosome 2 with the Genomic Association and Prediction Integrated tool (GAPIT) package in R [26]. Using markers on chromosome 2, rather than the whole set of 117,654 markers greatly reduce the computational speed for

the simulation. In the standard mixed linear model in (MLM) GWAS, the set up is

$$y = Wv + X\beta + Zu + e \tag{2.5}$$

where y is the vector of phenotypes, v and β are unknown fixed effects representing marker effects and non-marker effects, respectively; and u is a vector of size n (number of individuals) for unknown random polygenic effects having a distribution with mean of zero and covariance matrix of $G = 2K\sigma_a^2$, where K is the kinship (co-ancestry) matrix with element k_{ij} $(i, j = 1, 2, , n)$ and σ_a^2 is unknown genetic variance. In CMLM, individuals were grouped and the kinship (K) is replaced by the kinship among groups to reduce the computational speed [25]. Using the GAPIT package is much slower then the single-marker Fisher's exact test, therefore, simulation was run 100 times with subsample sizes from 30 to 100 with 10 individuals as increment.

If the significant SNP found in the CMLM model is within 5Mb distance from the marker S2_64198488, we consider it to be the correct marker. False positive markers were counted if there is any significant marker beyond the 5Mb range from S2_64198488 for each simulation.

Chapter 3

Results

3.1 Single SNP simulation

To simplify the case, we first simulated a total of $m = 100$ SNP markers with one SNP marker explaining 1%, 5% or 10% with $\beta_1 = 1$, and the other 99 markers considered to be 'noise' with $\beta_2 = \beta_3 = \ldots = \beta_{100} = 0$ (see formula (2.4)). Two allele frequencies at $p = 0.1$ and $p = 0.5$ were used for the simulation. For each of the six cases, we chose the number of extreme individuals (n_{sub}) ranging from a minimum of $n_s = 10$ to $n_l = 100$ with $s = 5$ increment. After sub-setting the samples, we assigned the top 50% of the response variables as 1, and bottom 50% as 0. Fisher's exact test was applied for each simulated data-set, and the P-value was extracted. For example, at allele frequency $p = 0.5$ and $R^2 = 0.05$ with $n_{sub} = 20$ subsample from each tail, the 2×2 table from Fisher's exact test is shown in Table 3.1, and the P-value to test for association in this table is 0.0002. Simulation for each case was replicated 1000 times. Power is calculated based on the proportion of P-values lower than a threshold (α level). The empirical power was plotted for each case with respect to the subsample sizes (Figures 3.1 and 3.2).

Figures 3.1 and 3.2 show that the simulation results with allele frequencies $p = 0.1$

Table 3.1: Example for Fisher's exact test with $p = 0.5$, $R^2 = 0.05$ and $n_{sub} = 20$ at each tail

	x_1	
y	0	1
0	18	2
1	6	14

Table 3.2:

	Allele Frequency	
R^2	0.1	0.5
0.01	> 1000	> 1000
0.05	106	90
0.10	39	31

and $p = 0.5$ respectively at α levels 10^{-5}, 10^{-6}, and 10^{-7}. The general trend is that higher thresholds require larger sample sizes for detecting significance, and lower variance explained by the SNP marker (R^2) also requires larger sample size. When the SNP marker explains a small amount of variance, for example 0.01, the extreme phenotype sampling method is not powerful to detect statistical significance; indeed the power to detect small effect SNP markers is low even for the complete sample (Table 3.2). The theoretical power of detecting SNP markers (Table 2.1) is relatively low at high alpha levels. When the effect of the SNP is moderate, explaining 5% of the total variance for example, a total of 106 and 90 extreme individuals are needed at each tail to reach 80% statistical power. The subsample number decreases to 39 and 31 at each tail when a SNP marker contributes 10% of the total variance (Table 3.2).

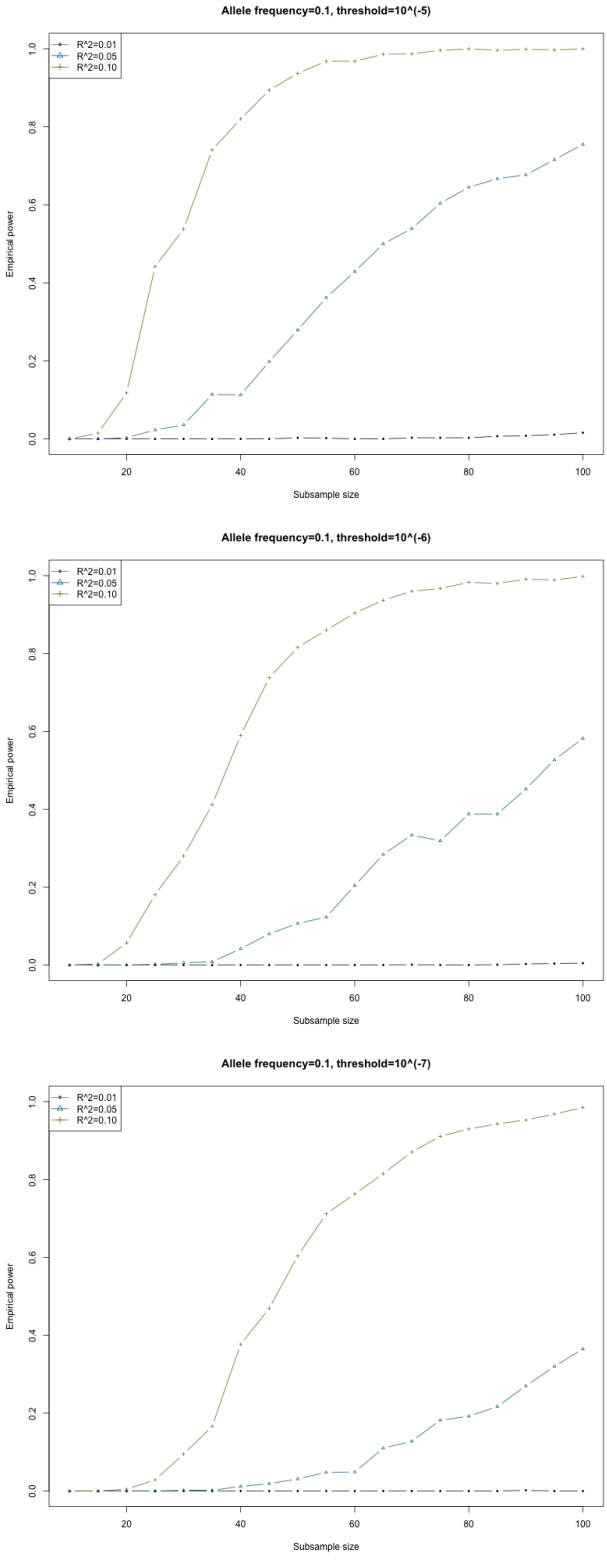

Figure 3.1: Single SNP simulation with $p = 0.1$ and significant thresholds at α levels 10^{-5}, 10^{-6}, and 10^{-7}

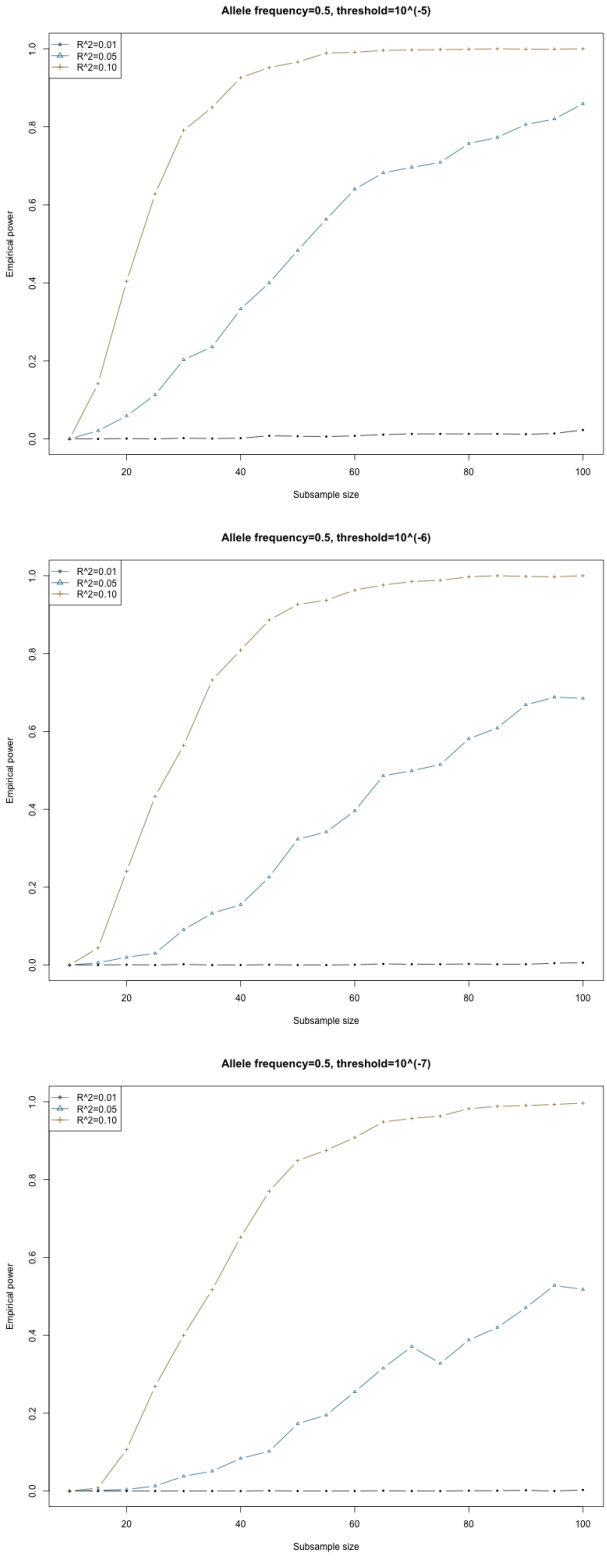

Figure 3.2: Single SNP simulation with $p = 0.5$ and significant thresholds α levels at 10^{-5}, 10^{-6}, and 10^{-7}

3.2 Multiple SNP scenario

3.2.1 Simple regression in the actual data-set

A previous study of an actual data set discovered a total of four chromosomal regions influencing plant height [20]. We selected a total of four SNP markers, S4_52707130, S6_42138323, S7_56824744 and S9_57236778 either at or near the peak SNP markers from the imputed data with different allele frequencies and variance explained (Table 2.2). We subset the data by selecting the individuals with extreme plant height after natural log transformation, and used the Fisher's exact test for each respective marker. Transformed P-value [-/log10(P-value)] was plotted with respective to sub-sample sizes (Figure 3.3). This result suggested that subsampling can detect two SNP markers, S6_42138323 and S9_57236778, requiring a total of 19 and 12 extreme individuals to be significant at alpha level of 10^{-5}. The other two SNP markers, S4_52707130 and S7_56824744 are difficult to detect with the subsampling strategy as they are of small effect than the other two SNP markers even in the complete data-set (Table 2.2).

3.2.2 Multiple SNP simulation

In multiple SNP simulation, we simulated a total of 100 SNP markers with four, x_1, x_2, x_3 and x_4 collectively explaining 40% of the total variance based on the actual data (details in Materials and Methods). The sampling method is the same with single SNP simulation. After each subsample was chosen based on extreme phenotypic values, we transformed the phenotypes to 0 and 1 for the lower 50% and higher 50% of data respectively, then we used the Fisher's exact test for each SNP marker for 1000 simulations. Empirical power against sample sizes was plotted. Figure 3.4 suggests that the sub-sampling strategy is more powerful when the single SNP marker explains more variance. The explanatory variable x_4, which accounted for 20% of total variance with an allele frequency $p = 0.40$, requires

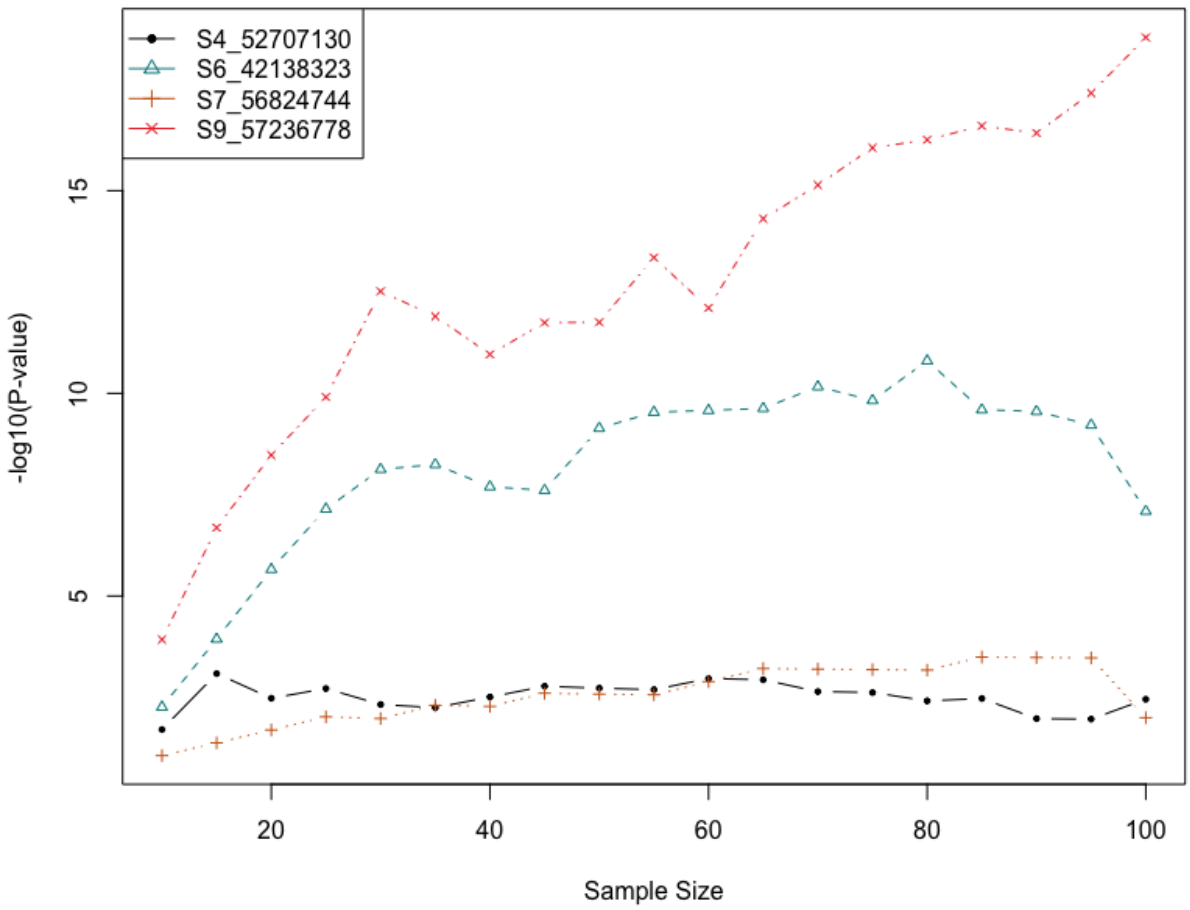

Figure 3.3: P-value against sub-sample sizes for simple regression of each SNP against sorghum plant height [20]

only 16 samples at each tail to reach 80% statistical power at threshold of 10^{-5} (Table 3.3). Two explanatory variables, x_1 and x_2, both explaining 5% of total variance, require 122 and 103 individuals at each tail to reach 80% statistical power, respectively. Detecting x_1 requires more individuals, possibly because of lower allele frequency (Table 3.3). Increasing the significant threshold requires more samples at each tail to reach 80% statistical power, as is shown in Figure 3.4.

The conclusion that we can detect two SNPs (S6_42138323 and S9_57236778) explaining relatively large variance with 19 and 12 individuals at each extreme tail in the actual data-set is consistent with simulation results (Table 3.3), which suggested 31 and 16 individuals to reach 80% statistical power. However, the subsampling strategy does not work well with S4_52707130 and S7_56824744. The simulation suggested that a total of 122 and 103 individuals at each tail were needed at the alpha level of 10^{-5}) (Table 3.2), while in the actual data-set, subsampling with Fisher's exact tests fail to reach such a low P-value.

Table 3.3: Samples needed at each tail to reach 80% power for each scenario at alpha level of 10^{-5} for multiple SNP simulation

	x_1	x_2	x_3	x_4
Variance	5%	5%	10%	20%
Allele freq	0.07	0.15	0.30	0.40
Sample size at each tail	122	103	31	16

3.3 Single SNP simulation with the actual data-set

3.3.1 simulation result

A trait was simulated based on one SNP marker on chromosome 2, S2_64198488, with allele frequency $p = 0.469$, and explaining 10% of the total variance using equation (2.5). A total of 15,222 SNP markers on chromosome 2 was used for the analysis. A mixed-model approach controlling for population structure for GWAS analysis [25] was used for each SNP marker

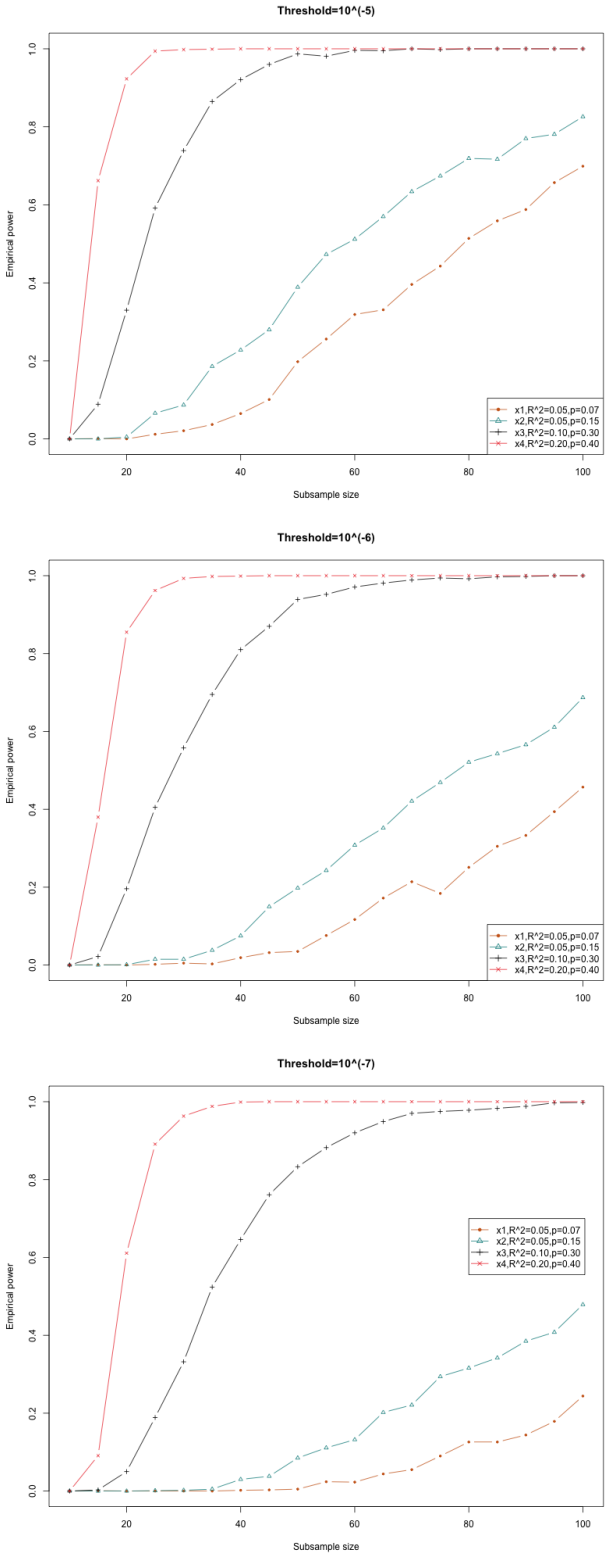

Figure 3.4: Simple regression for multiple SNP simulation result at significant thresholds $\alpha = 10^{-5}$, $\alpha = 10^{-6}$, and $\alpha = 10^{-7}$

in simulation for the power analysis with GAPIT [26]. The same setting of simulation also used ANOVA and Fisher's exact test for comparison.

The 100 simulation result suggested that the linear regression has the highest power for detecting SNP S2_64198488, while GAPIT has the lowest power (Figure 3.5), possibly because GAPIT controls for population structure, i.e., relationships between individuals. In addition, the statistical power increases with increased sub-sample size, also suggested by the previous analysis. While GAPIT has the lowest power to detect the causal marker, it also has the lowest false positive rate (Figure 3.6) among the three methods at α level of 10^{-5}. ANOVA has the largest false positive rate among the three methods.

3.3.2 Power analysis with varying initial population size

The previous analysis (Table 3.2) suggested that a total of 31 individuals are needed to reach 80% power to detect a SNP marker with allele frequency $p = 0.5$ and explaining 10% of the total variance at α level of 10^{-5} with Fisher's exact test, sampling from a population size of $n = 1000$. While in section 3.3.1, similar analysis with the Fisher's exact test suggested approximately 70 individuals are needed to detect the associated markers, sampling from a population size of $n = 354$.

We conducted the same simulation as in the section 2.2 with allele frequency $p = 0.5$ and $R^2 = 0.10$ but varying initial starting population size(n), from $n = 200$ to $n = 1200$. Figure 3.7 has suggested that with small initial population size, the power of extreme individual sampling methods is relatively low. More individuals from the extremes are needed to reach the same statistical power when the initial sample size is relatively small. This is probably because with large sample size, the population shows large dispersion so that the causal SNP is easier to detect. When the initial population size is above 800, differences in power detection related to population size are no longer obvious (Figure 3.7).

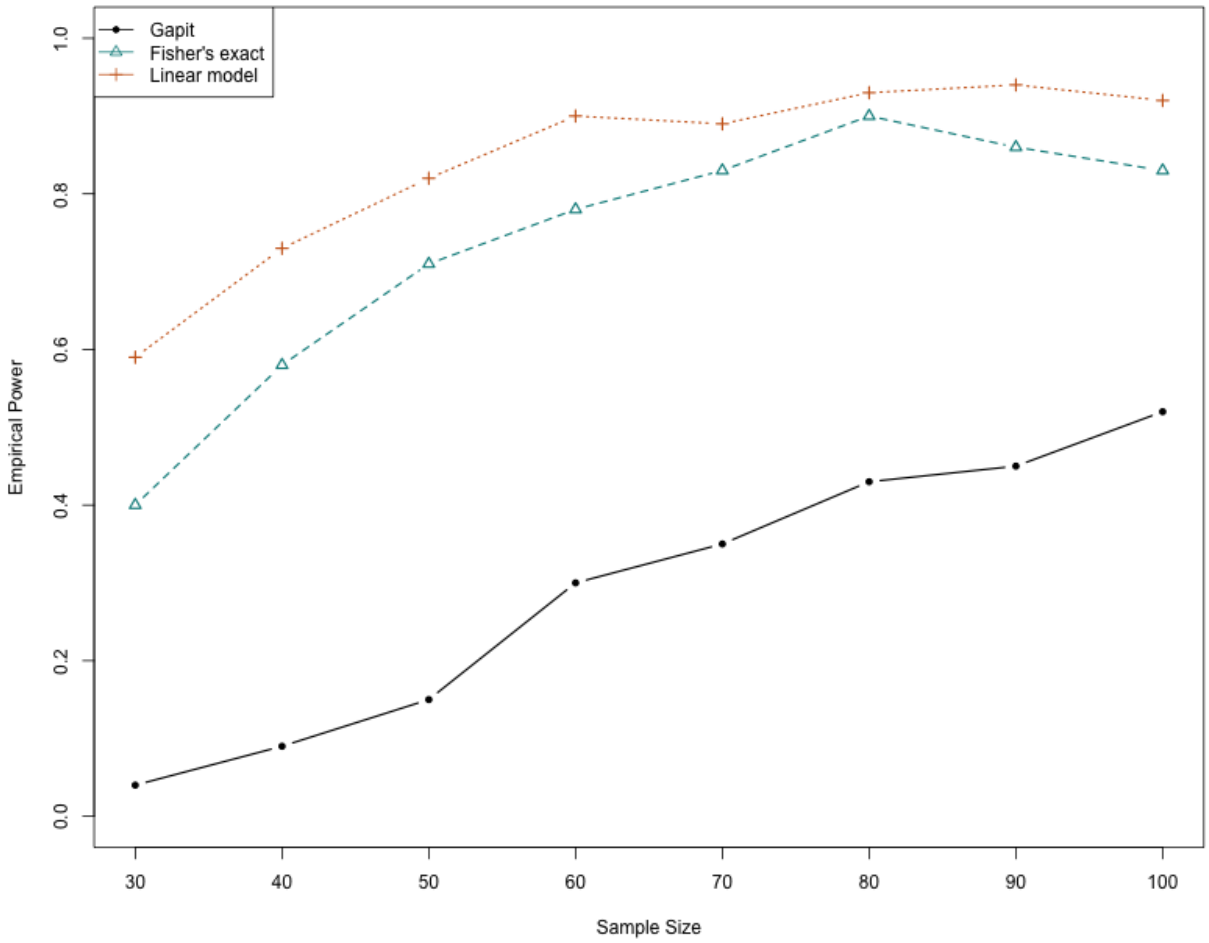

Figure 3.5: Power analysis comparing three different methods at alpha level of 10^{-5}

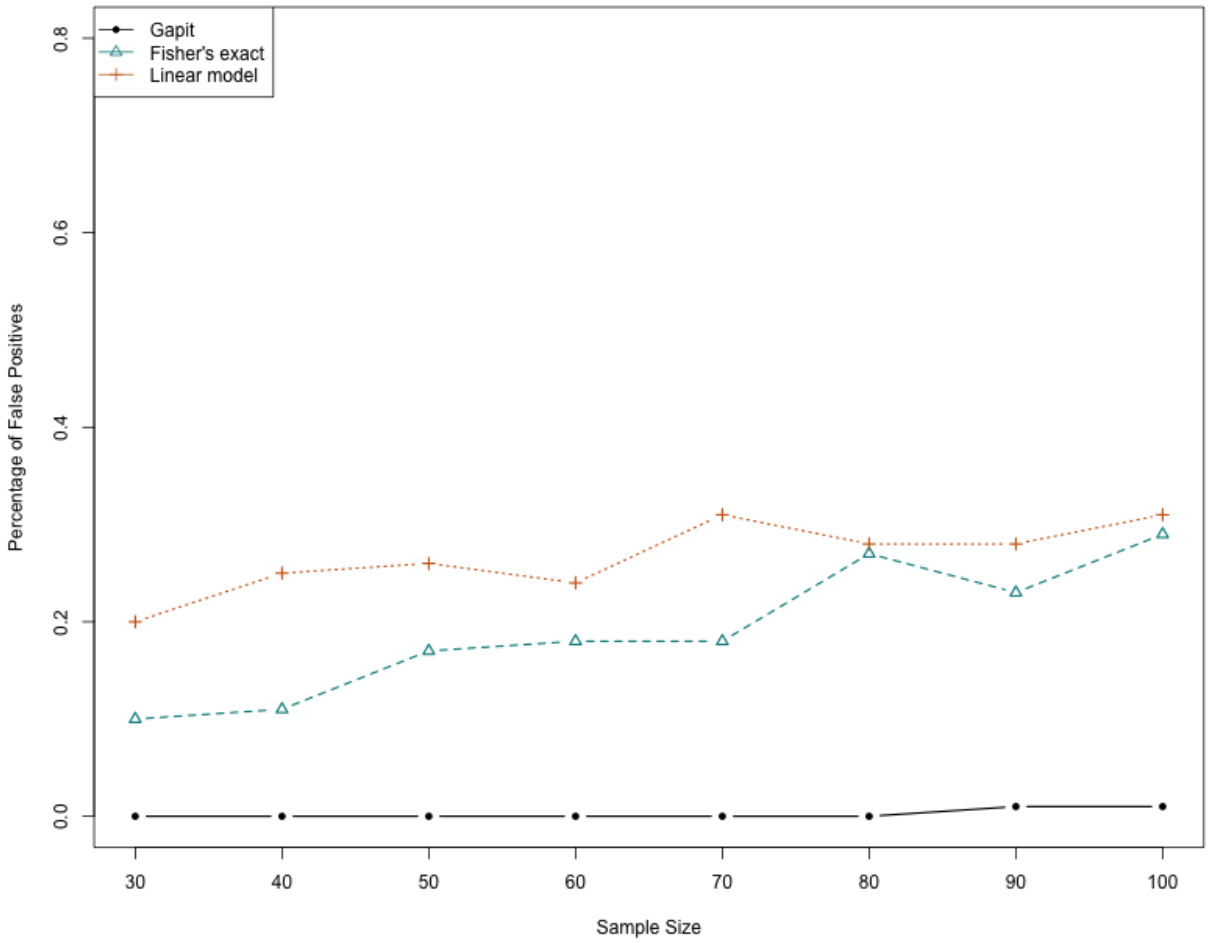

Figure 3.6: False positives rates for three different methods at α level of 10^{-5}

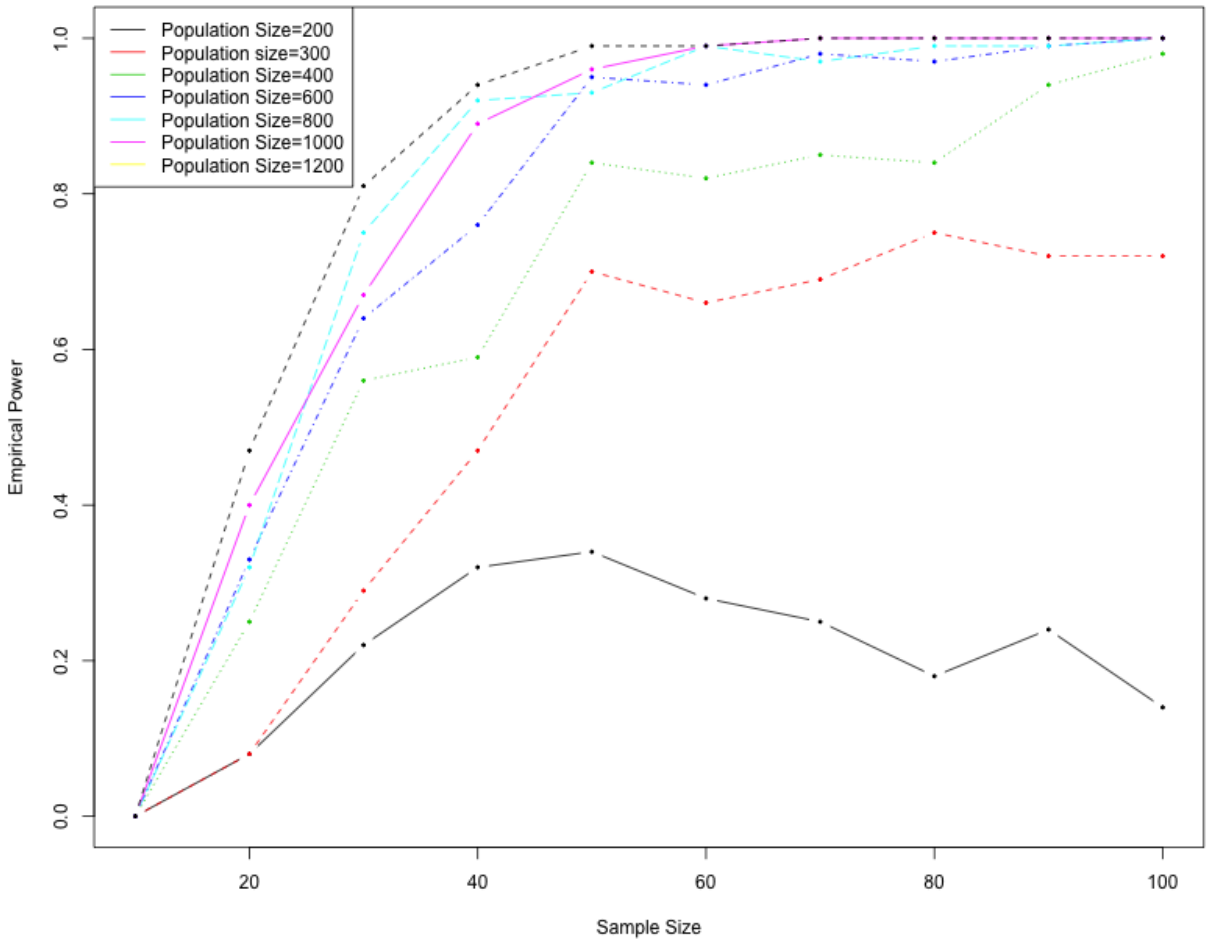

Figure 3.7: Empirical power against sub-sample size with different initial population sizes

Chapter 4

Conclusion

We have evaluated statistical power of selective genotyping strategies based on sampling extreme individuals in a GWAS design. This sampling method, though not as powerful as using the whole GWAS panel statistically, is potentially economical in real-world experimental designs to save genotyping costs for quantitative trait locus (QTL) detection and validation. We provided guidance on determining the minimum sample size required to reach 80% of statistical power (Table 3.2) both for single and multiple allele scenarios (Table 3.2 and Table 3.3) with different allele frequency (p) and variance explained (R^2) using simulation. We compared the power and false positive rates of three different statistical methods applied in the real GWAS panel in sorghum (Figure 3.5 and 3.6), and concluded that the popular package GAPIT has the lowest false positive rate by sacrificing detection power, while a linear model (ANOVA) has both the highest power and false positive rate. In addition, we discovered that the statistical power of detecting causal SNP marker in selective genotyping is also dependent on the initial population size, with larger initial population size conferring greater dispersion and increasing detection power.

References

[1] Chao Xing and Guan Xing. Power of selective genotyping in genome-wide association studies of quantitative traits. In *BMC proceedings*, volume 3, page S23. BioMed Central, 2009.

[2] Paul M Magwene, John H Willis, and John K Kelly. The statistics of bulk segregant analysis using next generation sequencing. *PLoS computational biology*, 7(11):e1002255, 2011.

[3] Eric S Lander and David Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199, 1989.

[4] R W Michelmore, I Paran, and RV Kesseli. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the national academy of sciences*, 88(21):9828–9832, 1991.

[5] GL Wang and AH Paterson. Assessment of dna pooling strategies for mapping of qtls. *Theoretical and Applied Genetics*, 88(3-4):355–361, 1994.

[6] Martin Trick, Nikolai Maria Adamski, Sarah G Mugford, Cong-Cong Jiang, Melanie Febrer, and Cristobal Uauy. Combining snp discovery from next-generation sequencing

data with bulked segregant analysis (bsa) to fine-map genes in polyploid wheat. *BMC Plant Biology*, 12(1):14, 2012.

[7] David L Hyten, James R Smith, Reid D Frederick, Mark L Tucker, Qijian Song, and Perry B Cregan. Bulked segregant analysis using the goldengate assay to locate the rpp3 locus that confers resistance to soybean rust in soybean. *Crop Science*, 49(1):265, 2009.

[8] Paul L Auer and Guillaume Lettre. Rare variant association studies: considerations, challenges and opportunities. *Genome medicine*, 7(1):16, 2015.

[9] Jinliang Yang, Haiying Jiang, Cheng-Ting Yeh, Jianming Yu, Jeffrey A Jeddeloh, Dan Nettleton, and Patrick S Schnable. Extreme-phenotype genome-wide association study (xp-gwas): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel. *The Plant Journal*, 84(3):587–596, 2015.

[10] Dalin Li, Juan Pablo Lewinger, William J Gauderman, Cassandra Elizabeth Murcray, and David Conti. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic epidemiology*, 35(8):790–799, 2011.

[11] Ian J Barnett, Seunggeun Lee, and Xihong Lin. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic epidemiology*, 37(2):142–151, 2013.

[12] BE Huang and Danyu Y Lin. Efficient association mapping of quantitative trait loci with selective genotyping. *The American Journal of Human Genetics*, 80(3):567–576, 2007.

[13] Or Zuk, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander. Searching

for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.

[14] Lin T Guey, Jasmina Kravic, Olle Melander, Noël P Burtt, Jason M Laramie, Valeriya Lyssenko, Anna Jonsson, Eero Lindholm, Tiinamaija Tuomi, Bo Isomaa, et al. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genetic epidemiology*, 35(4):236–246, 2011.

[15] Sofie Van Gestel, Jeanine J Houwing-Duistermaat, Rolf Adolfsson, Cornelia M van Duijn, and Christine Van Broeckhoven. Power of selective genotyping in genetic association analyses of quantitative traits. *Behavior genetics*, 30(2):141–146, 2000.

[16] Neil Risch and Heping Zhang. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, 268(5217):1584–1589, 1995.

[17] Michael A Gore, Jer-Ming Chia, Robert J Elshire, Qi Sun, Elhan S Ersoz, Bonnie L Hurwitz, Jason A Peiffer, Michael D McMullen, George S Grills, Jeffrey Ross-Ibarra, et al. A first-generation haplotype map of maize. *Science*, 326(5956):1115–1117, 2009.

[18] Xuehui Huang, Tao Sang, Qiang Zhao, Qi Feng, Yan Zhao, Canyang Li, Chuanrang Zhu, Tingting Lu, Zhiwu Zhang, Meng Li, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics*, 42(11):961, 2010.

[19] Geoffrey P Morris, Punna Ramu, Santosh P Deshpande, C Thomas Hash, Trushar Shah, Hari D Upadhyaya, Oscar Riera-Lizarazu, Patrick J Brown, Charlotte B Acharya, Sharon E Mitchell, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences*, 110(2):453–458, 2013.

[20] Dong Zhang, Wenqian Kong, Jon Robertson, Valorie H Goff, Ethan Epps, Alexandra Kerr, Gabriel Mills, Jay Cromwell, Yelena Lugin, Christine Phillips, et al. Genetic

analysis of inflorescence and plant height components in sorghum (panicoidae) and comparative genetics with rice (oryzoidae). *BMC plant biology*, 15(1):107, 2015.

[21] Andrew H Paterson, John E Bowers, Remy Bruggmann, Inna Dubchak, Jane Grimwood, Heidrun Gundlach, Georg Haberer, Uffe Hellsten, Therese Mitros, Alexander Poliakov, et al. The sorghum bicolor genome and the diversification of grasses. *Nature*, 457(7229):551, 2009.

[22] Yurii S Aulchenko, Stephan Ripke, Aaron Isaacs, and Cornelia M Van Duijn. Genabel: an r library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296, 2007.

[23] Holger Schwender and with a contribution of Arno Fritsch. *scrime: Analysis of High-Dimensional Categorical Data such as SNP Data*, 2013. R package version 1.3.3.

[24] JR Quinby, RE Karper, et al. Inheritance of height in sorghum. *Inheritance of height in sorghum.*, 1953.

[25] Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, et al. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4):355, 2010.

[26] Alexander E. Lipka, Feng Tian, Qishan Wang, Jason Peiffer, Meng Li, Peter J. Bradbury, Michael A. Gore, Edward S. Buckler, and Zhiwu Zhang. Gapit: genome association and prediction integrated tool. *Bioinformatics*, 28(18):2397–2399, 2012.