GWAS Project Report
Version 5.0

Prepared for
Dr. Yan Yan, Dr. Nisha Puthiyedth,
Department of Computing Science, TRU
Comp 4910

Prepared by
Nuoyi Zhang T00592163
Ziqing Wang T00055783

May 6th, 2021

Table of contents

1 INTRODUCTION

GAWS is a "main technology for identifying the association between genetic variation and traits or diseases." (Genome-Wide Association Studies (GWAS), 2021) This technique can be used to "understand how genes cause diseases and develop better prevention and treatment strategies." (Genome-Wide Association Studies (GWAS), 2021) SNPs are single nucleotide polymorphisms. The basic data in GWAS data analysis includes genotype data and phenotype data. “Genotype is the complete set of the genetic material of an organism. However, the genotype is usually used to refer to a single gene or a group of genes, such as the genotype of eye color.” (Genotype - Wikipedia, 2021) “In genetics, phenotype is a collection of observable characteristics or traits of an organism.” (Phenotype - Wikipedia, 2021)

Although GWAS has been used for a long time in the past, GWAS analysis often faces challenges. One of them is that a small sample contains a large amount of SNP data. Typically, millions of SNPs have thousands or hundreds of samples. The objective is to find true causal of SNPs in a huge dataset, typically only hundreds of them. "Data sparse", therefore, the feature selection method provides the feasibility of reducing SNP. However, the feature selection method is not fully applicable to large genomes and GWAS problems. So, we plan to use the Least Absolute Shrinkage and Selection Operator (LASSO) method as the basis. Two approaches within LASSO method will be evaluated BIGLASSO and AUTALASSO.

2 METHODS

AUTALASSO "proposed an automatic adaptive lasso based on the Alternating Direction Multiplier Method (ADMM) optimization algorithm." (Waldmann et al., 2019) It "provides higher prediction accuracy” (Waldmann et al., 2019)and ” compared with the ordinary lasso, the lasso has lower prediction error, GWAS capability with additive and dominance effects at the same time." (Waldmann et al., 2019)

BIGLASSO uses memory-mapped files to store a large amount of data on disk. Due to memory limitations, the existing R software package cannot adapt to lasso-type models to be adapt to genetics and genomics, so BIGLASSO only uses data during model fitting when necessary. Read into memory, so it can seamlessly process out-of-core calculations. Moreover, it is equipped with newly proposed and more effective feature screening rules, which can greatly accelerate the calculation speed.

The entire project process is to run the method using the same input data for the two methods, AUTOLASSO and BIGLASSO, and draw the results from each method respectively. The output files of the two methods are first compared and analyzed with the PLINK results. As a basic analysis software in GWAS, PLINK is used as the standard tool in the project to compare the output results of the two methods. To ensure the accuracy of the two methods. After comparing with PLINK, we performed a horizontal comparison of the two methods, in order to explore the relationship between the two methods.

3 EXPERIMENTAL RESULTS AND DISCUSSIONS

First of all, there are four testing datasets in this project. The first two files the file test data that comes with the two methods. The 3rd input file is TESSL's tutorial data file. The last input file was downloaded from the easyGWAS website. It is the Arabidopsis thaliana（AtPolyDB）data

（https://easygwas.ethz.ch/down/1/）

There are three files in the Arabidopsis thaliana dataset: genotype.ped, genotype.map, phenotypes.pheno.

The ped file and map file store the genotype SNP data; the pheno file stores the phenotype data.

About ped files and map files:

The two-file PED/MAP format often contain both family-based and regular genotype data popularized by PLINK and can be imported into Array Studio. The "ped" file format refers to the widely-used format for linkage pedigree data. Check out the information at the PLINK website on the "ped" file format" (Joseph, 2017).

In short, the PED format will start with six fields in each row:

Family ID ('FID')

Within-family ID ('IID'; cannot be '0')

Within-family ID of father ('0' if father isn't in dataset)

Within-family ID of mother ('0' if mother isn't in dataset)

Sex code ('1' = male, '2' = female, '0' = unknown)

Phenotype value ('1' = control, '2' = case, '-9'/'0'/non-numeric = missing data if case/control)

followed by 2*(number of variants) columns, calling each allele in the order of the .map file. (Joseph, 2017)

The phenotype.pheno contains all the phenotype data of Arabidopsis thaliana. We will select the required phenotypes for testing.

3.1 BIGLASSO

3.1.1 Experimental Data

Data name: colon.Rdata

Introduction: The test data that comes with BIGLASSO. This data file contains gene expression data of 62 samples (40 tumor samples, 22 normal samples) from colon-cancer patients analyzed with an Affymetrix oligonucleotide Hum6000 array (Zeng et al., 2021).

Download address: https://github.com/YaohuiZeng/biglasso

Data size: 380 KB

Data format: Rdata

Data name: genotype.map, genotype.ped, phenotypes.pheno

Introduction: Arabidopsis thaliana dataset with 1307 samples.

Download address: https://easygwas.ethz.ch/data/public/dataset/view/1/

Data size: genotype.map: 5.92 MB, genotype.ped: 1.04 GB, phenotypes.pheno: 583 KB

Data format: .map, .ped, .pheno. These formats are text formats suitable for PLINK

3.1.2 Replication Results

This is the code I used to reproduce the results of the

BIGLASSO research:

```
library(ncvreg)
library(Matrix)
library(bigmemory)
library(biglasso)
library("writexl")

data(colon)
X <- colon$X
y <- colon$y
dim(X)
X[1:5, 1:5]
y
X.bm <- as.big.matrix(X)
str(X.bm)
dim(X.bm)
X.bm[1:5, 1:5]
fit <- biglasso(X.bm, y, family = "binomial")

coefs <- as.matrix(coef(fit, lambda = 0.0522))
coefs[coefs != 0, ]
predict(fit, lambda = 0.0522, type = "nvars")
predict(fit, lambda = 0.0522,   type = "vars")
```

The results are the same as those shown in the BIGLASSO paper (Zeng & Breheny, 2018):

Hsa.8147	Hsa.36689	Hsa.42949	Hsa.22762	Hsa.692.2	Hsa.31801
249	377	617	639	765	1024
Hsa.3016	Hsa.5392	Hsa.1832	Hsa.12241	Hsa.44244	Hsa.2928
1325	1346	1423	1482	1504	1582
Hsa.41159	Hsa.33268	Hsa.6814	Hsa.1660		
1641	1644	1772	1870		

To further test the accuracy of BIGLASSO in gwas, we used the Arabidopsis thaliana dataset:

PLINK code:

PLINK --file test_file –recodeA

PLINK --file genotype --allow-no-sex --assoc –out PLINK_output --adjust

3.1.3 Code Documents

get_input_with_Emco5.py:
https://github.com/ZiqingWang774649749/GWAS/blob/main/Biglasso/get_input_with_Emco5.py

get_input_with_FT10.py:
https://github.com/ZiqingWang774649749/GWAS/blob/main/Biglasso/get_input_with_FT10.py

biglasso_test.R:
https://github.com/ZiqingWang774649749/GWAS/blob/main/Biglasso/biglasso_test.R

new_genotype.py:
https://github.com/ZiqingWang774649749/GWAS/blob/main/Biglasso/new_genotype.py

3.1.4 Process

DNA is two parallel polynucleotide chains entwined with each other to form a double helix structure. The unit that makes up this chain is called deoxyribonucleotide. There are 4 types according to different bases, namely A (adenine), T (thymine), C (cytosine), G (guanine). The bases on the same DNA follow a certain pairing principle, that is, A and T are paired , C and G are paired.

First, we need to convert the ATGC in the ped file into 0,1,2 format, because BIGLASSO can only recognize SNP data in 0, 1, and 2 format, it cannot recognize ATGC format. I use PLINK code: PLINK --file test_file --recodeA to convert ped and map files into p1.raw file.

From Figure 2 we can see that the format of the first six columns in p1.raw is the same as that of genotype.ped in Figure 1, except that the sixth column has changed from 0 to -9, and both 0 and -9 here represent lack of data. Starting from the seventh column The SNP data becomes 012 format. The missing part of the genotype value becomes NA, the major of snp becomes 0, the minor of snp becomes 2, and the heterozygosity becomes 1.

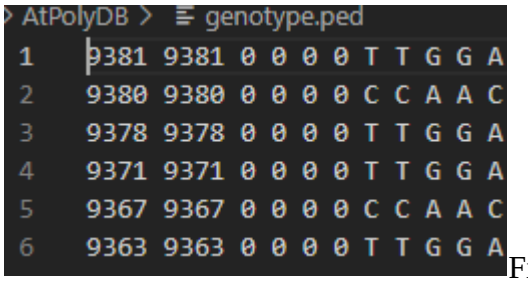

Figure 1

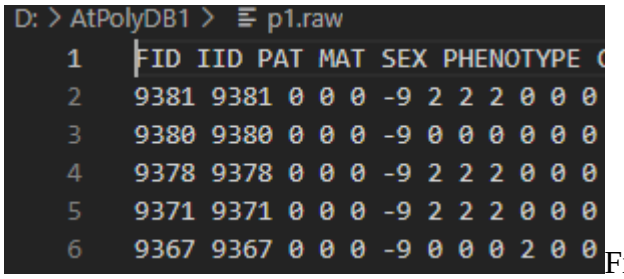

Figure 2

Then I need to find the required phenotypes from phenotypes.pheno and add it to p1.raw.

For phenotype in 1,0 format:

Use get_input_with_Emco5.py to delete the first six columns of p1.raw and find snp Emco5 from phenotype.pheno and add it to p1.raw as the first column, and delete the line containing Nan in the snp Emco5. The final output is re_p1.raw.

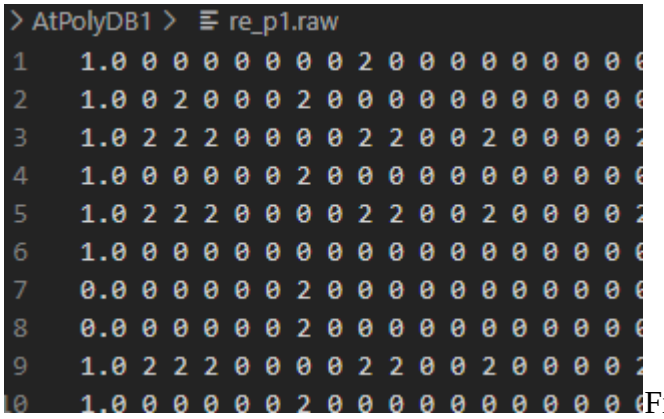

Figure 3

We can see that in Figure 3, the first column of re_p1.raw is completely composed of 0 and 1. re_p1.raw will be used as the input of BIGLASSO.

For phenotype whose format is not 1,0 format:

use get_input_with_FT10.py to delete the first six columns of p1.raw and find snp FT10 from phenotype.pheno and add it to p1.raw as the first column, and delete the line containing Nan in the snp Emco5. Then I sorted according to FT10 and changed the first half of FT10 to 0 and the second half to 1. The final output is biglasso_input_with_FT10.raw.

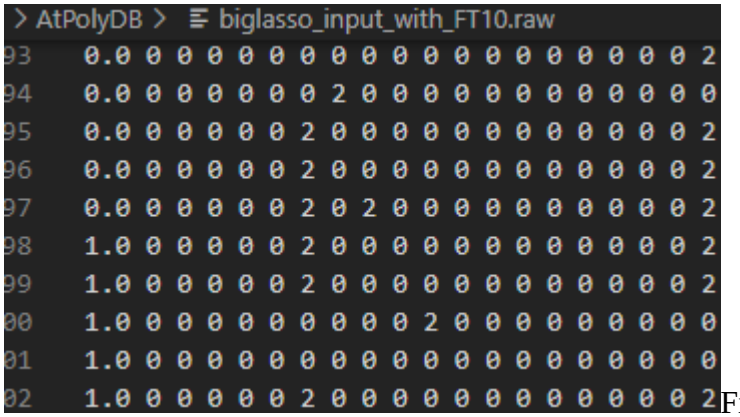

Figure 4

We can see that in Figure 4, the first column of biglasso_input_with_FT10.raw is completely composed of 0 and 1. re_p1.raw will be used as the input of BIGLASSO.

Second, use BIGLASSO to run the input file just prepared, and here is my R code:biglasso_test.R.

Copy the result to output.txt.

(Intercept)	V2320	V7183	V52175	V57068	V62384	V77100	V96655
3.2084334930	-0.6863507850	-0.1335295719	-0.0304324895	-0.7850377564	-0.3876059939	-0.1455658922	-0.1719689036
V107838	V125676	V141714	V156789	V164947	V186781	V191171	V198309
-0.1398064072	-0.1956704167	-0.0248685383	-0.1280789300	-0.1040785866	-0.0562255356	-0.0005252101	-0.0583875338
V201633	V203391	V204761	V206419	V209653			
-0.4789444976	-0.3883797848	-0.4305592400	-0.2291353508	-0.0994738479			

Figure 5

In Figure 5, the V+ number represents the order of the snp in genotype.map, we can use it later to find the name of the corresponding snp.

Third, I use new_genotype.py to replace snp Emco5/FT10 in the sixth column of genotype.ped and got genotype2.ped and ran it with PLINK.

PLINK code:PLINK --file genotype_file --allow-no-sex --assoc --out PLINK_output_file –adjust

Compare the data output.txt obtained from BIGLASSO with PLINK_output.qassoc.adjusted

This file is sorted by significance value rather than genomic location, the most significant results being at the top (Purcell et al., 2014).

3.1.5 Result Documents

plink_Emco5_output.qassoc.adjusted:
https://github.com/ZiqingWang774649749/GWAS/blob/main/Biglasso/plink_Emco5_output.qassoc.adjusted

plink_FT10_output.qassoc.adjusted:
https://github.com/ZiqingWang774649749/GWAS/blob/main/Biglasso/plink_FT10_output.qassoc.adjusted

Result_Documents_Emco5.txt:
https://github.com/ZiqingWang774649749/GWAS/blob/main/Biglasso/Result_Documents_Emco5.txt

Result_Documents_FT10.txt:
https://github.com/ZiqingWang774649749/GWAS/blob/main/Biglasso/Result_Documents_FT10.txt

I use get_geno_name.py to get the name of the snp in the output.txt file and compare it with the ranking in plink_output.qassoc.adjusted, and finally write the ranking in the last column by hand.

In Result_Documents_Emco5.txt, the third column is the name of snp, and the last column is their ranking in plink_output.qassoc.adjusted:

2320 Chr1 Chr1_1430179 0 1430179 4 3

7183 Chr1 Chr1_4266002 0 4266002 12 11

52175 Chr2 Chr2_155225 0 155225 10 9

57068 Chr2 Chr2_2990590 0 2990590 2 1

62384 Chr2 Chr2_6949821 0 6949821 63 62

77100 Chr2 Chr2_17434080 0 17434080 78 77

96655 Chr3 Chr3_9028198 0 9028198 28 27

107838 Chr3 Chr3_14993958 0 14993958 19 18

125676 Chr4 Chr4_783577 0 783577 17 16

141714 Chr4 Chr4_8380917 0 8380917 67 66

156789 Chr4 Chr4_16076083 0 16076083 145 144

164947 Chr5 Chr5_2232651 0 2232651 7 6

186781 Chr5 Chr5_13945171 0 13945171 21 20

191171 Chr5 Chr5_15982859 0 15982859 120 119

198309 Chr5 Chr5_18567388 0 18567388 86 85

201633 Chr5 Chr5_20112353 0 20112353 3 1

203391 Chr5 Chr5_21041671 0 21041671 18 17

204761 Chr5 Chr5_21697730 0 21697730 6 5

206419 Chr5 Chr5_22625167 0 22625167 9 8

209653 Chr5 Chr5_24456951 0 24456951 331 330

The ranking in this file looks very good. Many SNPs produced by BIGLASSO have high rankings.

Result_Documents_FT10.txt, The result of the comparison is not ideal, only a few SNPs have a high ranking.

Maybe converting the phenotype to 0 and 1 is not a good choice.

The time value in Emco5 is 0.83s, in FT10 is 1.55s.

3.1.6 Machine Specs

Operating System: Windows 10 Home 64-bit (10.0, Build 18363)

Processor: Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz (4 CPUs), ~3.3GHz

Memory: 8192MB RAM

3.2 AUTALASSO

3.2.1 Experimental Data

QTLMAS2010ny012.csv

Introduction: This data was included in AUTALASSO package. It was used to test the AUTALASSO method running.There are 3226 SNPs in the data file.

Data size: 61285KB
Data format: CSV
Download address: patwa67/AUTALASSO (github.com)

mdp_genotype.vcf

Introduction: TASSEL tutorial dataset was second-try input file.

Data size: 3514KB
data format:VCF
download address: Tassel (bitbucket.io)

genotype.ped

genotype.map

phenotypes.pheno

Introduction: The final dataset we decided to use for our project.Arabidopsis thaliana dataset with 1307 samples, 214051 SNPs.

Data size:1092854KB, 6070KB, 584KB
Data format: PED, MAP, PHENO.
Download addresss: easyGWAS - Public Data (ethz.ch)

3.2.2 Replication Results

The results obtained in the AUTOLASSO test run in this project are not compared with the original output results. The reason is that after the test is run, it is confirmed that the AUTOLASSO runs successfully and the test run is stopped, and the results of glmnet are obtained again without following the steps of the original AUTOLASSO.

3.2.3 Code Document

AUTALASSO run file:

GWAS-COMP4910/ATUALASSO.ipynb at main · NotA1Lmight/GWAS-COMP4910 (github.com)

Data conversion file:

GWAS-COMP4910/AUTALASSO_FT10_RAW_TO_CSV.py at main · NotA1Lmight/GWAS-COMP4910 (github.com)

3.2.4 Process

Step.1

Input: I am using the same input name p1.raw from Ziqing uploaded. In case, our different input files lead us to wrong direction.

Step.2

data file conversion. AUTALASSO cannot handle the raw file as input. So, I am using the AUTALASSO_FT10_RAW_TO_CSV.py to change the raw file to csv file. Because the phenotype for first time was using Emco5 as first column in input file. That is binary value. So, there are all zero value fill up the output file.

Step. 3

Change the phenotype from Emco5 to FT10.That means the input file need change also. p1.raw is still working before adding phenotype FT10 to the first column. Addition, change raw file to csv in one code file AUTALASSO_FT10_RAW_TO_CSV.py.

Step 4

Using AUTALASSO run the input. For output, there are four output file. Three of four are lambda value. One is the result after running AUTALASSO code.

Step 5.

Because the output values are in one column. There are more than 630000 rows value. After organization, which means putting three rows value in one row and deleting the zero value rows. In this step, there are 189 result. After doing maximum and minimum top 10, 3 and bottom 10, 3.All result in one file outbeta.xlsx(Figure 6,7,8 showing below).

			SNPs address	SNPs-Name
0	0	-0.024407557	3146	Chr1_1964828
-0.717592393	0	0	4522	Chr1_2817007
0	-0.046896526	0	5446	Chr1_3314951

Figure 6

Record of qassoc file								p-value
1	Chr1_1964828	1964828	194	-2.574	1.939	0.009091	-1.327	0.186	0.186
1	Chr1_2817007	2817007	194	-4.37	3.692	0.007245	-1.184	0.238	0.238
1	Chr1_3314951	3314951	194	1.695	1.98	0.003799	0.8557	0.3932	0.3932

Figure 7

Record of adjusted file									adjusted-Ranking
1	Chr1_1964828	0.186	0.4709	1	1	1	1	0.3964	1	100447
1	Chr1_2817007	0.238	0.5201	1	1	1	1	0.4573	1	111410
1	Chr1_3314951	0.3932	0.6419	1	1	1	1	0.6081	1	138417

Figure 8

Running time 181.505267 seconds (4.56 M allocations: 40.129 GiB, 1.66% gc time)

Figure 6 shows the real SNPs name by rows number. Figure 7 tells the p-value which means compared with PLINK value in plink_FT10_output.qassoc file. Figure 8 means the ranking of this SNP with the plink_FT10_output.qassoc.adjusted file from PLINK.

3.2.5 Result Document

PLINK output file:

GWAS-COMP4910/plink_FT10_output.qassoc at main · NotA1Lmight/GWAS-COMP4910 (github.com)

GWAS-COMP4910/plink_FT10_output.qassoc.adjusted at main · NotA1Lmight/GWAS-COMP4910 (github.com)

AUTALASSO output file:

GWAS-COMP4910/Final_outbeta.zip at main · NotA1Lmight/GWAS-COMP4910 (github.com)

3.2.6 Machine Specs

Operating System: Windows 10 pro 64-bit (10.0, Build 19042) (19041.vb_release.191206-1406)

Processor: AMD Ryzen 5 3500X 6-Core Processor (6 CPUs), ~3.6GHz

Memory: 16384MB RAM

3.3 Comparison Results

The last two files were compared by comparing the SNP names, and the same SNP was not found.

And because AUTALASSO cannot use the phenotype of binary value as feature selection. BIGLASSO is just the opposite. BIGLASSO can only run phenotype input files with binary values. First of all, these two methods are very complementary in feature selection. The output of AUTALASSO needs to be sorted again and there is no label, so it will feel confusing for people who are new to contact. Relatively speaking, BIGLASSO will append a number to the output to indicate the position of the SNP in the original input file. So BIGLASSO was recommended in this project experiment.

Three different data sets were used in this experiment. The first two datasets are included in the method. For the project, two dataset only run with testing. So, first two datasets are not suit for project running. Second dataset has lots of Nan value. After deleting Nan value, there are few data which means dataset has no enough sample and SNPs for project running The third is a dataset downloaded from the easyGWAS website. 1307 samples, 214051 SNP, 107 phenotypes. Guaranteed, the final result is sufficient. It satisfies the choice of binary phenotype and non-binary phenotype. So as the final goal of the data set, we got the desired results and data.

4 CONCLUSIONS AND FUTURE WORK

For this project, two methods were run with R and julia respectively. Get the corresponding result. In the process, the project encountered many difficulties. For example, data format conversion, test results comparison, and data selection. But in the end, the research on the project concluded that the two methods of AUTOLASSO and BIGLASSO have different solutions for different phenotype feature.

In future plans, the project will continue to test different phenotypes based on the last dataset to compare horizontally, whether the same SNPs will appear in the output file. Or use the new GWAS analysis tool as the standard value for comparison and exploration again. Whether there will be different output results for different phenotype and standard values.

5 REFERENCE

Genome.gov. 2021. *Genome-Wide Association Studies (GWAS)*. [online] Available at: <https://can01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.genome.gov%2Fgenetics-glossary%2FGenome-Wide-Association-Studies%23%3A%7E%3Atext%3DA%2520genome-wide%2520association%2520study%2520%2528GWAS%2529%2520is%2520an%2520approach%2Cused%2520to%2520predict%2520the%2520presence%2520of%2520a%2520disease.&data=03%7C01%7C%7Ca5bf097ffc0449d19474320d8418bd0a%7Ceb1c9d1ae6e8409787febb01690935b7%7C0%7C0%7C637553789759999999%7CGood%7CV0FDfHsiViI6IjAuMC4wMDAwIiwiUCI6IiIsIkFOIjoiIiwiV1QiOjR9&sdata=JaFwe8D4fGYdDZNV00p3QY3A1m128hl3Dw%2BwQwcGdd4%3D&reserved=0> [Accessed 6 May 2021].

Joseph. (2017, May 30). *SNP Data: PED file + Map file*. Array Suite Wiki. http://www.arrayserver.com/wiki/index.php?title=SNP_Data%3A_PED_file_%2B_Map_file

Purcell, S., Center for Human Genetic Research, Massachusetts General Hospital, Broad Institute of Harvard, & MIT. (2014, May 15). *purcell lab*. PLINK: Whole genome data analysis toolset. http://zzz.bwh.harvard.edu/PLINK/

Waldmann, P., Ferenčaković, M., Mészáros, G. *et al.* AUTALASSO: an automatic adaptive LASSO for genome-wide prediction. *BMC Bioinformatics* **20,** 167 (2019). https://doi.org/10.1186/s12859-019-2743-3

Waldmann, P., Ferenčaković, M., Mészáros, G., Khayatzadeh, N., Curik, I., & Sölkner, J. (2019). AUTALASSO: an automatic adaptive LASSO for genome-wide prediction. *BMC bioinformatics*, *20*(1), 167. https://doi.org/10.1186/s12859-019-2743-3

Wikipedia contributors. (2021, April 24). Phenotype. In *Wikipedia, The Free Encyclopedia*. Retrieved 07:34, April 30, 2021, from https://en.wikipedia.org/w/index.php?title=Phenotype&oldid=1019669652

Wikipedia contributors. (2021, January 11). Genotype. In *Wikipedia, The Free Encyclopedia*. Retrieved 07:34, April 30, 2021, from https://en.wikipedia.org/w/index.php?title=Genotype&oldid=999778856

Zeng, Y., & Breheny, P. (2018, Mar 11). *The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R*. The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R. https://arxiv.org/pdf/1701.05936v2

Zeng, Y., and Breheny, P. (2017). The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R. arXiv preprint arXiv:1701.05936. URL https://arxiv.org/abs/1701.05936.

Zeng, Y., Wang, C., & Breheny, P. (2021, Apr 8). *biglasso: Extend Lasso Model Fitting to Big Data in R*. GitHub. https://github.com/YaohuiZeng/biglasso

6 ACKNOWLEDGEMENT

Special thanks to Dr. Kevin O'Neil.

Special thanks to Canshield.