Genetic association between genetic variation and flowering traits of *Miscanthus* grown at Alabama

Yongli Zhao
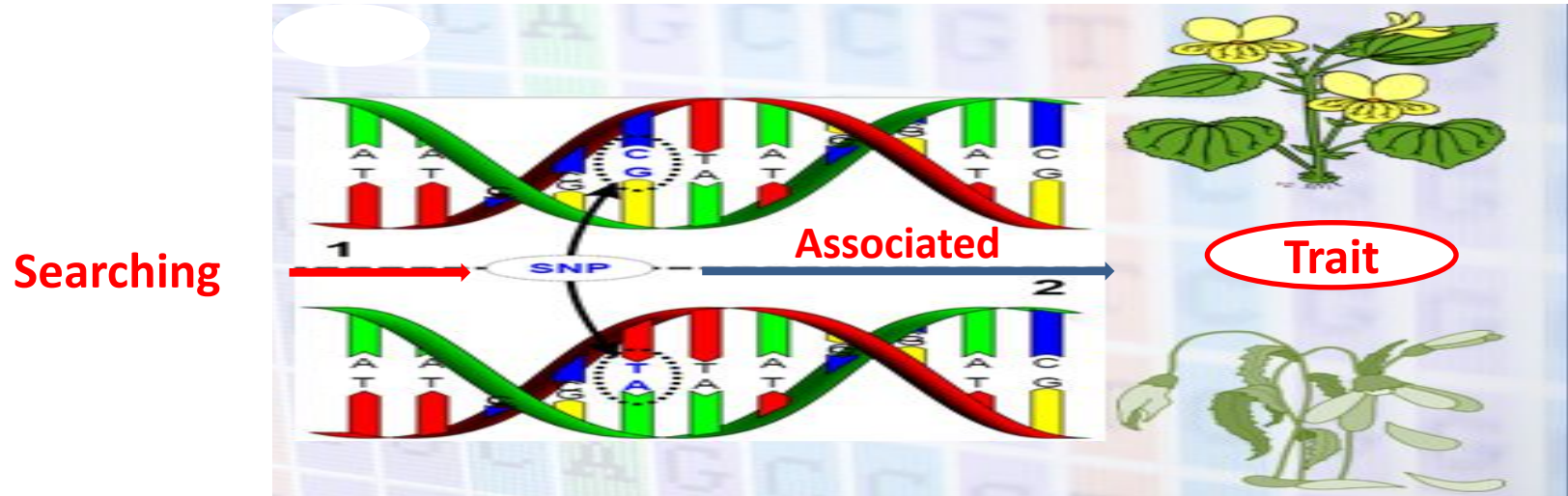
Ph.D, Biology, Tuskegee University
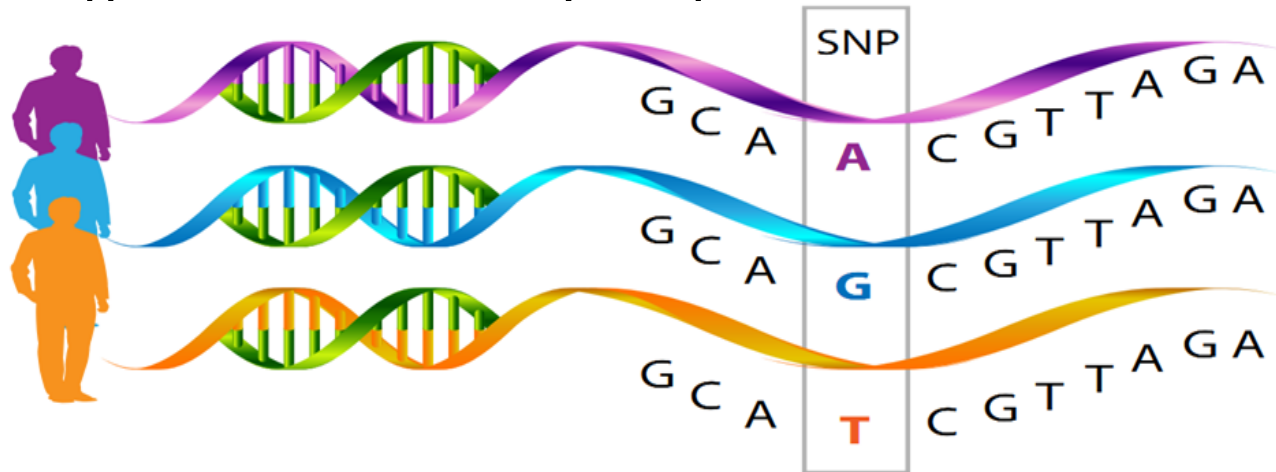
M.S. Probability and Statistics, Auburn University

Project

Goal: Finding the SNPs associated with flowering time trait.

GWAS: Genome-wide association study

SNP: Single Nucleotide Polymorphism

Data

Genotype data: SNPs

Names of individuals   **ID of the SNPs**

	A	B	C	D	E	F	G	H	I	J	K	L	M
Taxa	GFSSContig1	GFSSContig1	GFSSContig1	GFSSContig1	GFSSContig1	GFSSContig1	GFSSContig1	GFSSContig1	GFSSContig1	GFSSContig2	GFSSContig3	GFSSContig3
CANE9233-U	0	2	2	2	2	2	2	2	0	0	2	2
DC-2010-001	0	0	2	2	2	0	2	1	0	2	2	2.22E-16
DC-2010-001	0	1	2	2	2	0	2	2	0	1	2	1
KY-2009-001	0	2	2	2	2	0	2	0	0	1	2	1
KY-2009-001	0	1	2	2	1	1	2	1	0	1	2	1
KY-2009-001	0	1	2	2	2	0	2	0	0	2	2	2.22E-16
KY-2009-001	0	1	2	2	2	0	2	0	0	2	2	2.22E-16
NC-2010-001	0	2	2	2	1	0	2	1	0	0	2	1
NC-2010-001	0	2	2	2	1	1	2	2	0	0	2	2

Homozygotes are denoted by "0" and "2".
Heterozygotes are denoted by "1".

                                                              Imputed values

Genotype: ~46177 imputed SNPs developed by Dr. Lindsay Clark in UIUC.

https://raw.githubusercontent.com/yzhao4651/Miscanthus/master/data/subgenomrMLMMflo.csv

3

Data

Phenotype data: flowering time traits

Taxa	fprind	fprinW	fprinGW	fprinM
CANE9233-US47-0011	NA	NA	NA	NA
DC-2010-001-A	0.00445564	0.000572	0.00147371	-0.0084716
DC-2010-001-E	-0.0531224	-0.006452	-0.0052058	-0.0205294
KY-2009-001-C	0.06402954	0.00973598	0.00451221	0.00861331
KY-2009-001-D	-0.1257318	-0.0175624	-0.0173858	-0.0398475
KY-2009-001-B20-b	NA	NA	NA	NA
KY-2009-001-B20-e	0.52097233	0.06760393	0.06182684	0.0925469
NC-2010-001-B	0.22689188	0.02981267	0.0246261	0.04751059
NC-2010-001-009	0.33832821	0.04047861	0.03550158	0.07922877
NC-2010-001-010	-0.46976	-0.0636878	-0.0544341	-0.0593234
NC-2010-001-5-B18-a	-0.0619673	-0.0071505	-0.0075921	-0.0295163
NC-2010-001-5-B18-b	NA	NA	NA	NA
NC-2010-001-B44-e	0.155846	0.019176	0.01592664	0.02446856
NC-2010-001-5-C	0.52569055	0.06911147	0.05843188	0.09585178
NC-2010-002-001	0.59723241	0.08075154	0.06887779	0.11982047
NC-2010-002-002	NA	NA	NA	NA
NC-2010-003-D	0.40805454	0.05234427	0.04580888	0.0734315
NC-2010-003-E	-0.0692233	-0.0135194	-0.0115406	-0.0308912
NC-2010-003-001	NA	NA	NA	NA
NC-2010-003-B50-a	NA	NA	NA	NA

Missing value

N: 169 individuals for two years data

P: 4 variables

fprind: flowering time labeled with day.

fprinW: flowering time labeled with week.

fprinGW: flowering time labeled with week group

fprinM: flowering time labeled with Month.

https://github.com/yzhao4651/Miscanthus/blob/master/data/alltraits.csv

Experimental Approach

Data analysis: GWAS analysis

➢ **Multivariate methods (Multiple loci analysis)**

FarmCPU: k model (Liu et al. 2016)

https://github.com/yzhao4651/Miscanthus/blob/master/5%20Mis-imputedSNP-data%20incubator.R
https://github.com/yzhao4651/Miscanthus/blob/master/6%20GAPIT-and-FarmCPU-data%20incubator.R
https://github.com/yzhao4651/Miscanthus/blob/master/11%20FLOPC-data%20incubator.R

Result

Results

	SNP	Chromosome	Position	FarmCPU.fprind	FarmCPU.fprinW	FarmCPU.fprinGW	FarmCPU.fprinM
1	NsiI.TP620354	1	2563	0.48273189771	0.598694248	0.83573708	0.756836048
2	PstI.TP453248	1	3988	0.96149735691	0.036023208	0.58411336	0.193042555
3	GFSSContig17137_1860	1	35412	0.51134766731	0.609006034	0.86362874	0.326293876
4	GFSSContig17137_940	1	36536	0.92718949257	0.939163609	0.94023747	0.691244417
5	PstI.TP1112203	1	63590	0.88743429167	0.632239253	0.34340129	0.866978131
6	PstI.TP506588	1	80675	0.83735748683	0.641834738	0.36848171	0.325262208
7	PstI.TP989644	1	80828	0.62315130131	0.588820130	0.53115794	0.166047271
8	PstI.TP312105	1	81356	0.84317341887	0.874104786	0.60109384	0.332714741
9	NsiI.TP481742	1	154583	0.68820728601	0.879428023	0.37466130	0.320735860
10	PstI.TP863331	1	161269	0.96103133836	0.700892654	0.83625183	0.949334242

N: 36088 observations: each one is the SNPs name.

P: 7 variables.

SNP: the gene marker

Chromosome: the location of the gene marker

Position: the location in each Chromosome

FarmCPU.fprind: the P. Value from FarmCPU method with flowering time labeled with day

FarmCPU.fprinW: the P. Value from FarmCPU method with flowering time labeled with week

FarmCPU.fprinGW: the P. Value from FarmCPU method with flowering time labeled with week group

FarmCPU.fprinM: the P. Value from FarmCPU method with flowering time labeled with Month

Result
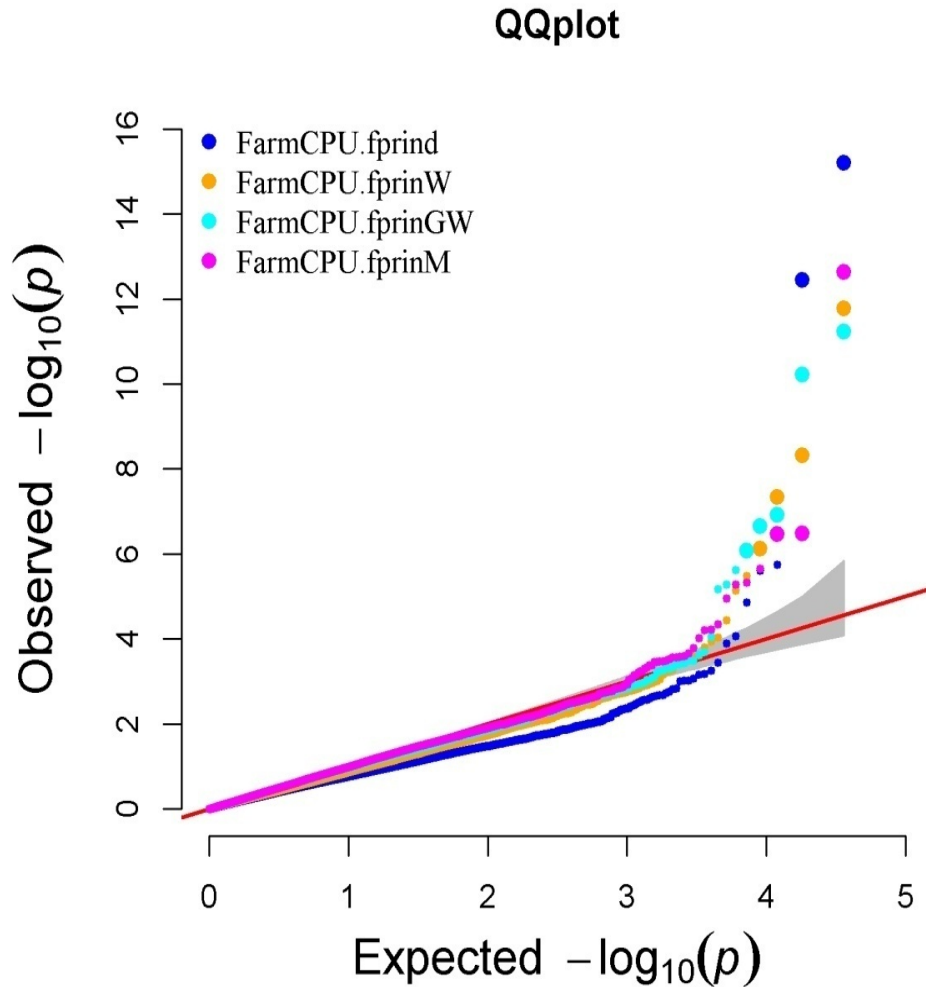
QQplot

The quantile-quantile (QQ) –plot:
a useful tool for assessing how well the model used in GWAS accounts for population structure and cryptic relatedness.
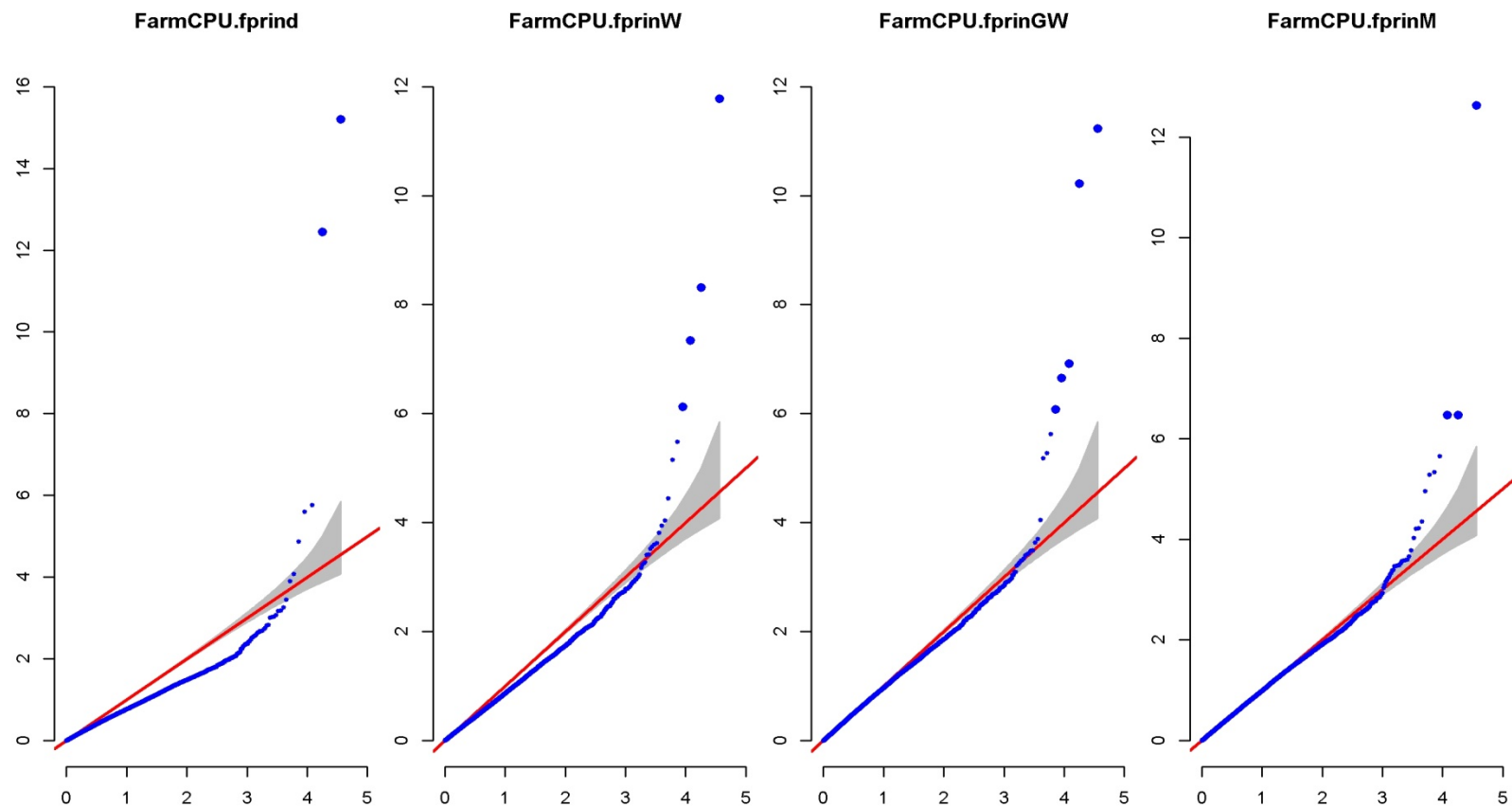- Y: -log10(*P*-values from the models)
  X: -log10(expected p-value of H0)
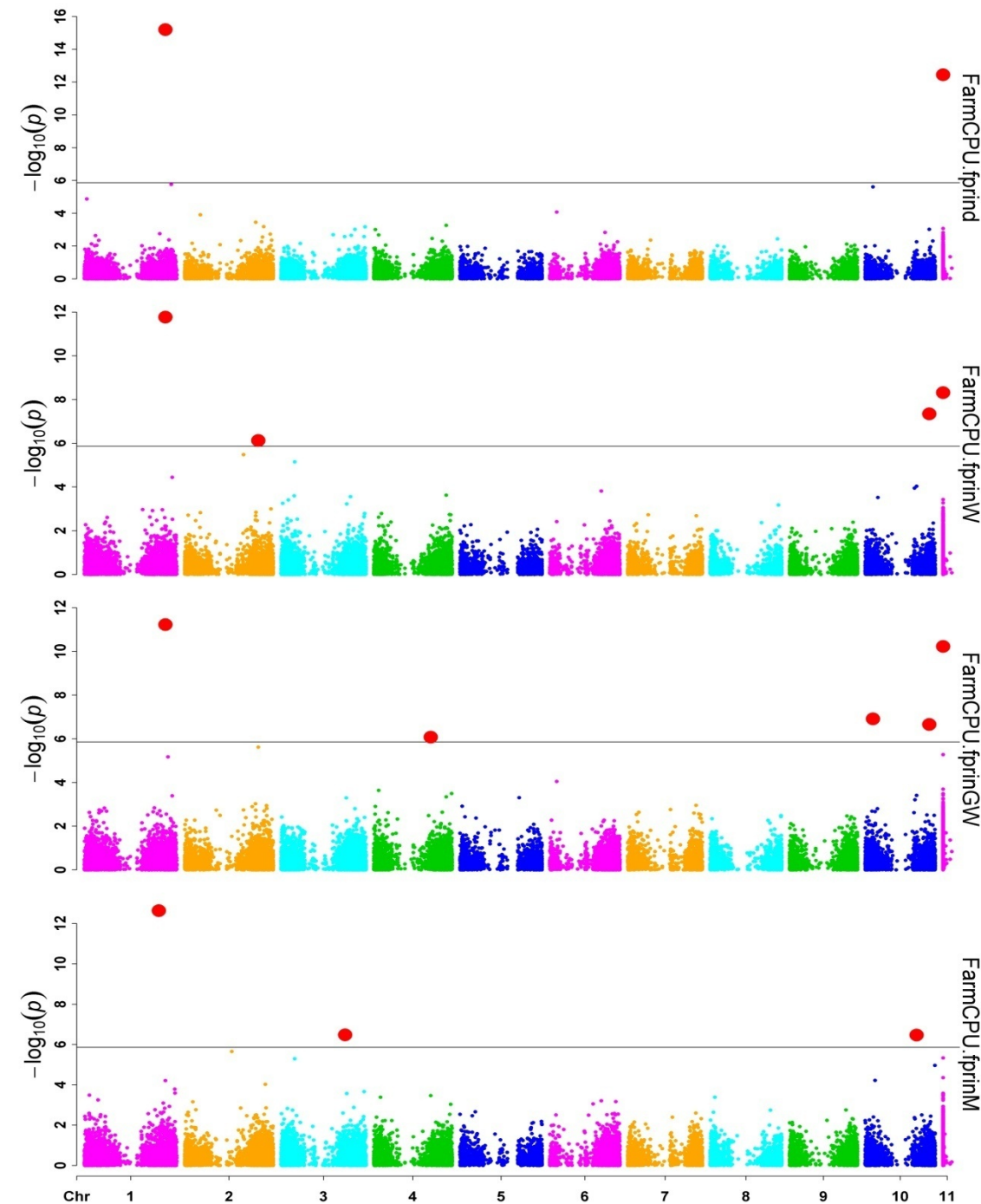   H0: no association

The majority of the points should lie on the diagonal line.

The SNPs on the upper right section of the graph deviate from the diagonal. These SNPs are most likely associated with the trait under study.

Deviations from the diagonal suggest the presence of spurious associations due to population structure and cryptic relatedness that the GWAS model does not sufficiently account for these spurious associations.

https://github.com/yzhao4651/Miscanthus/blob/master/Allimages/FarmCPU.all/Multracks.QQplot.FarmCPU.fprind.FarmCPU.fprinW.FarmCPU.fprinGW.FarmCPU.fprinM.jpg

Result (QQ-plot)

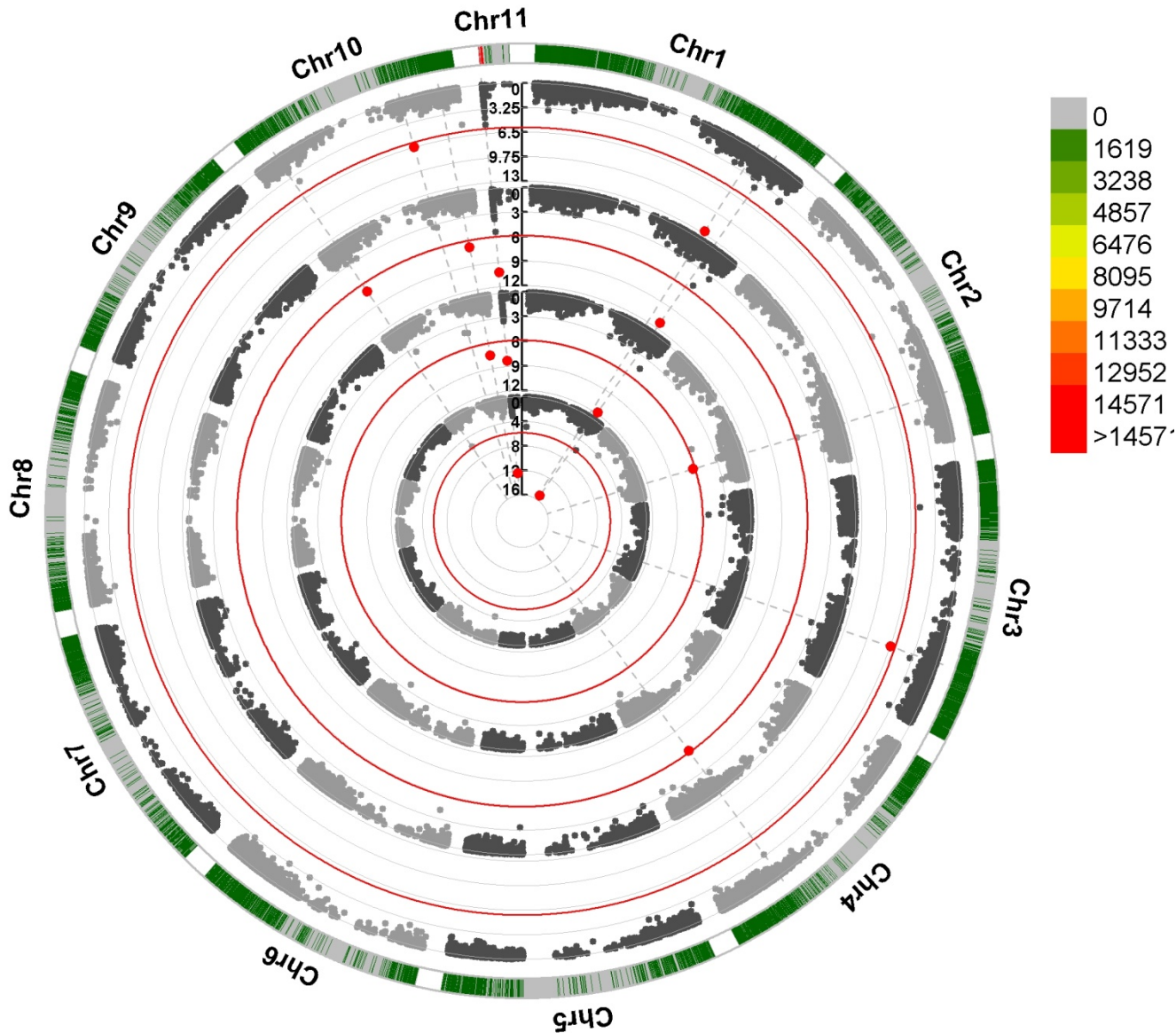Results (Manhattan plot)

A scatter plot that summarizes GWAS results

X-axis: the genomic position of the SNPs in the genome.

Y-axis: -log 10(*P*-values)

SNPs with stronger associations with the trait will have a larger Y-coordinate value.
(Red dots)

https://github.com/yzhao4651/Miscanthus/blob/master/Allimages/FarmCPU.all/Multracks.Rectangular-Manhattan.FarmCPU.fprind.FarmCPU.fprinW.FarmCPU.fprinGW.FarmCPU.fprinM.jpg

Result

Expectation: from Data incubator

1) labeling each of SNPs in the QQ-plots and Manhattan plots

2) Which statistics can be used to the best measure method for flowering time traits.

https://github.com/yzhao4651/Miscanthus/blob/master/11%20FLOPC-data%20incubator.R

https://github.com/yzhao4651/Miscanthus/blob/master/Allimages/FarmCPU.p.less.0.05.csv

Reference

1. Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PloS Genet 12(2): e1005767. doi:10.1371/journal.pgen.1005767
2. https://github.com/YinLiLin/R-CMplot