



Pre-breeding strategies for obtaining new resilient and added value berries

Genomic tools for berry: An introductory walk-through of a genomic association study

Bordeaux, Feb 14th 2024



GWAS

What is it?

GWAS stands for Genome-Wide Association Study. It is a type of study in genetics that **aims to identify genetic variations associated with a particular trait** or disease. In a GWAS, researchers analyze the genomes of a large number of individuals to identify common genetic variants that are more frequently found in individuals with the trait or disease of interest compared to those without.

What is it used for?

GWAS scans the entire genome, examining these variations to identify associations between specific genetic markers and the trait or disease being studied. These studies have been particularly useful in identifying **genetic factors associated with complex diseases** such as diabetes, heart disease, and various psychiatric disorders.



What do we need to perform a GWAS?

Phenotypic data (preferably good/robust data).

A genetic map or the genome sequence of your target organism

Markers that can be related to the genetic/genomic resources available.

Appropriate software

GWAS can be/is often performed as a two-step analysis:

1. Obtain phenotype data
 1. Field experiments
 2. Surveys
 3. Etc
2. Do the GWAS analyses

Which Software?

GAPIT



Genomic Association and Prediction Integrated Tool

Zhiwu Zhang Laboratory

WASHINGTON STATE
UNIVERSITY



Genomics Proteomics Bioinformatics 19 (2021) 629–640



ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



APPLICATION NOTE

GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction

Jiabo Wang^{1,2,*}, Zhiwu Zhang^{2,*}



GAPIT input data:

- Y Phenotype
- KI Kinship matrix
- CV Covariate variables
- G Genotype data in hapmap format
- GD Genotype data in numeric format
- GM Genotype map for numeric format.

What do we need to do association study with GAPIT?

Y = Phenotype data:

Taxa	EarHT	dpoll	EarDia
811	59.5	NaN	NaN
4226	65.5	59.5	32.21933
4722	81.13	71.5	32.421
33-16	64.75	64.5	NaN
38-11	92.25	68.5	37.897
A188	27.5	62	31.419
A214N	65	69	32.006
A239	47.88	61	36.064
A272	35.63	70	NaN
A441-5	53.5	67.5	35.008

GD = Genotype data:

taxa	PZB00859.1	PZA01271.1	PZA03613.2	PZA03613.1
33-16	2	0	0	2
38-11	2	2	0	2
4226	2	0	0	2
4722	2	2	0	2
A188	0	0	0	2

The markers in the 'GD' and in the 'GM' files **need** to be in the same order!

GM = Genetic map:

Name	Chromosome	Position
PZB00859.1	1	157104
PZA01271.1	1	1947984
PZA03613.2	1	2914066
PZA03613.1	1	2914171
PZA03614.2	1	2915078

...

This walk-through:

Phenotype data: Polyphenol content (pelargonidin-3-*O*-malonylglucoside) in strawberry lines.

Genotype data: Single nucleotide markers obtained with the iStraw35 array.

Genome: *F. vesca* v4.0

Software: GAPIT v3 under the R/Rstudio umbrella.

Davik et al. *Horticulture Research* (2020)7:125
<https://doi.org/10.1038/s41438-020-00347-4>

Horticulture Research
www.nature.com/hortres

ARTICLE

Open Access

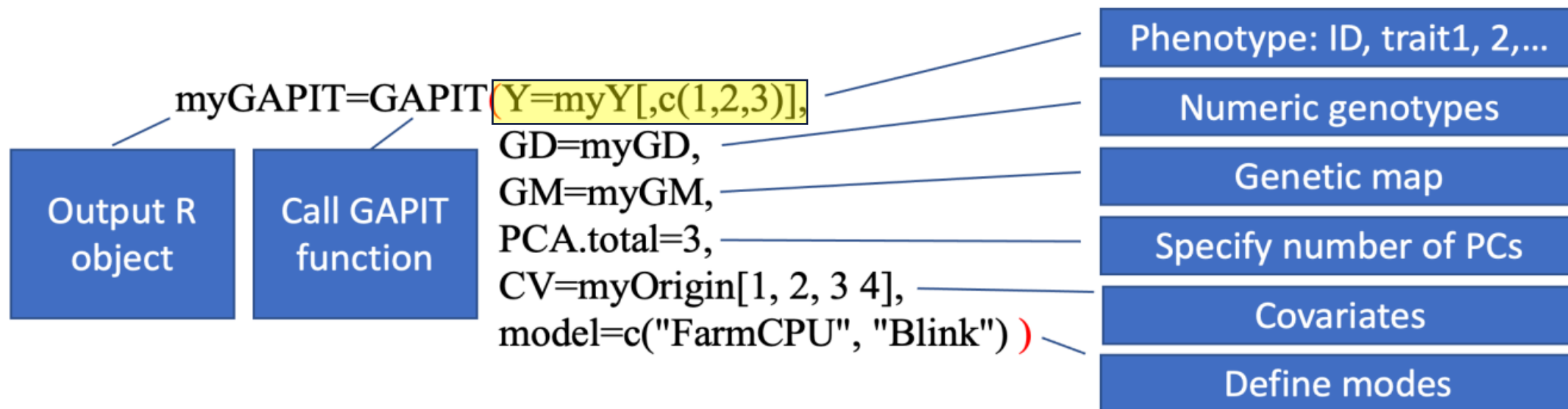
Major-effect candidate genes identified in cultivated strawberry (*Fragaria* × *ananassa* Duch.) for ellagic acid deoxyhexoside and pelargonidin-3-*O*-malonylglucoside biosynthesis, key polyphenolic compounds

Jahn Davik¹, Kjersti Aaby², Matteo Buti³, Muath Alsheikh^{4,5}, Nada Šurbanovski⁶, Stefan Martens⁷, Dag Røen⁴ and Daniel James Sargent⁸

05.03.2024



The basic GAPIT call:



Taxa	EarHT	dpoll	EarDia
811	59.5	NaN	NaN
4226	65.5	59.5	32.21933
4722	81.13	71.5	32.421
33-16	64.75	64.5	NaN
38-11	92.25	68.5	37.897
A188	27.5	62	31.419
A214N	65	69	32.006
A239	47.88	61	36.064
A272	35.63	70	NaN
A441-5	53.5	67.5	35.008

A closer look at our input data:

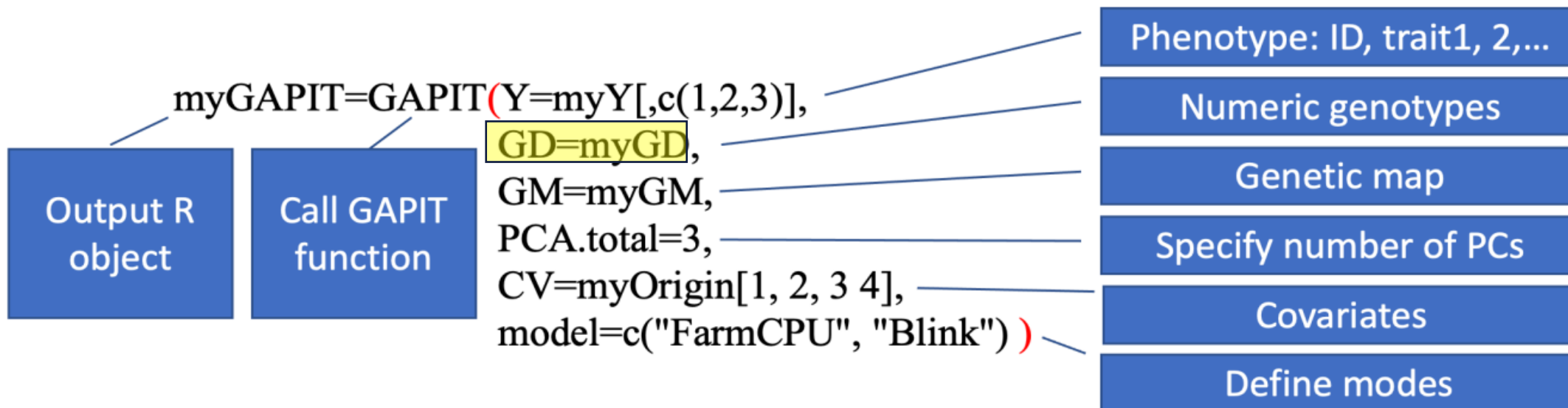
Phenotype data aka **Y** in previous slide.

```
> head(myY)
  taxa Pg_3_malglu
1 GT108_1 2.40e+00
2 GT108_10 2.75e+00
3 GT108_11 5.35e+00
4 GT108_12 1.00e-02
5 GT108_13 1.90e+00
6 GT108_14 1.33e-15
```

Pg_3_malglu == pelargonidin-3-*O*-malonylglucoside
(an anthocyanin)

Quantified by HPLC-DAD-MS of ripe fruit from three
mapping populations. Three replicates from each
strawberry line was used.

The basic GAPIT call:



taxa	PZB00859.1	PZA01271.1	PZA03613.2	PZA03613.1
33-16	2	0	0	2
38-11	2	2	0	2
4226	2	0	0	2
4722	2	2	0	2
A188	0	0	0	2

Need to filter the markers:

- Remove duplicates
- Minor allele frequency
- Call rate
- LD pruning?

Get myGD to the right format.

taxa	PZB00859.1	PZA01271.1	PZA03613.2	PZA03613.1
33-16	2	0	0	2
38-11	2	2	0	2
4226	2	0	0	2
4722	2	2	0	2
A188	0	0	0	2

Numeric genotypes aka **GD**

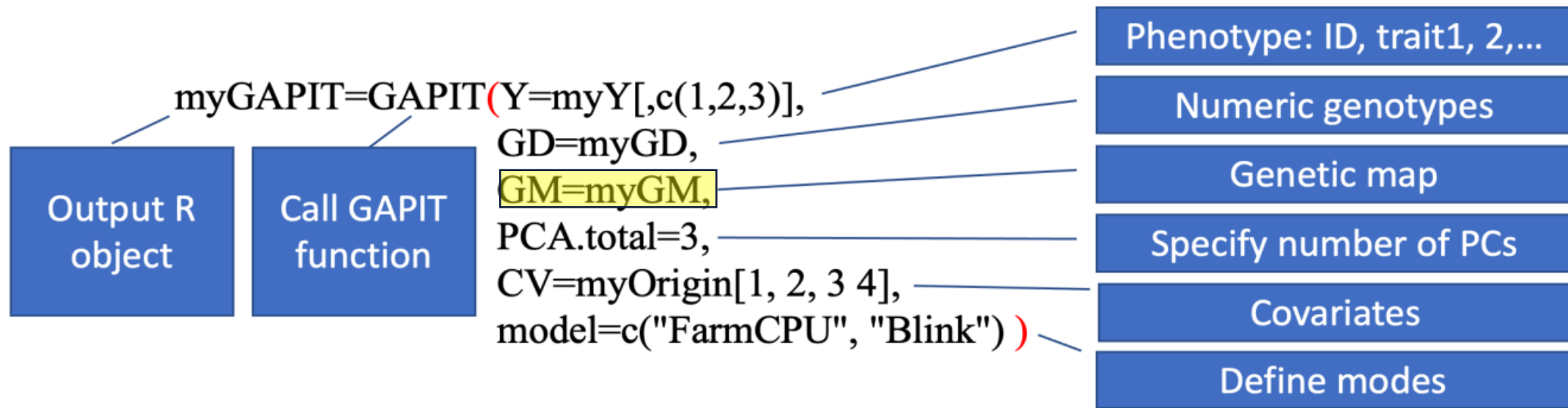
Taxa	AX-123356920	AX-123356921	AX-123356928	AX-123356929	AX-123356948	AX-123356977
GT108_1	0	2	1	2	1	0
GT108_2	0	0	2	2	2	2
GT108_3	0	2	1	2	2	0
GT108_4	1	1	2	2	2	1
GT108_5	1	1	1	2	2	1

2 == homozygote reference allele
 1 == heterozygote
 0 == homozygote alternative allele

Genotype data filtered on minor allele frequency
 Non-informative markers removed
 Markers on identical positions removed.

taxa	PZB00859.1	PZA01271.1	PZA03613.2	PZA03613.1
33-16	2	0	0	2
38-11	2	2	0	2
4226	2	0	0	2
4722	2	2	0	2
A188	0	0	0	2

The basic GAPIT call:

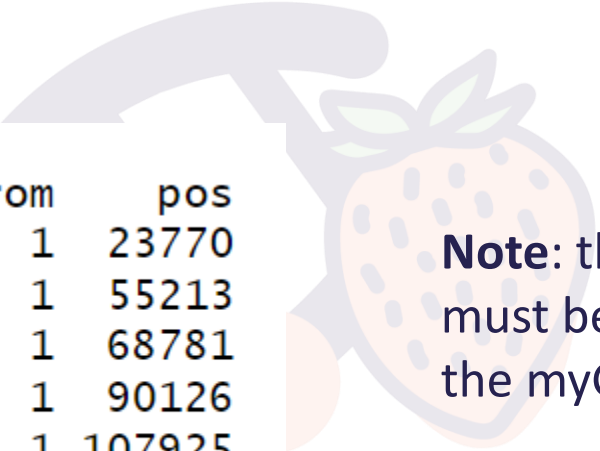


Name	Chromosome	Position
PZB00859.1	1	157104
PZA01271.1	1	1947984
PZA03613.2	1	2914066
PZA03613.1	1	2914171
PZA03614.2	1	2915078

...

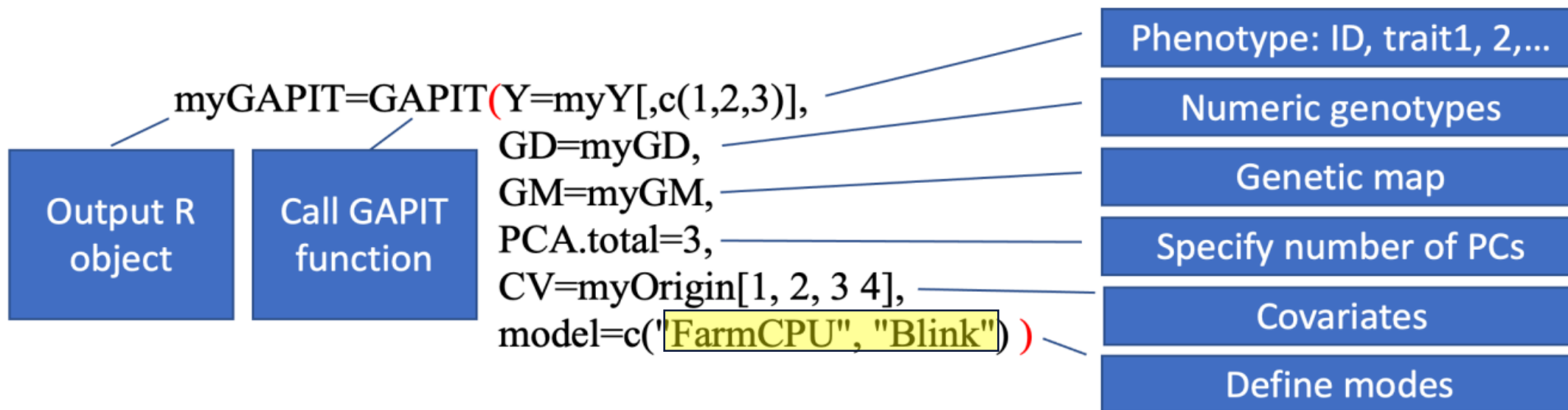
The map aka **GM**

```
> head(myGM)
      probe chrom  pos
4015 AX-123365097   1 23770
11988 AX-166510963   1 55213
11984 AX-166510957   1 68781
17723 AX-166520353   1 90126
11975 AX-166510934   1 107925
6631  AX-166502996   1 115335
> |
```

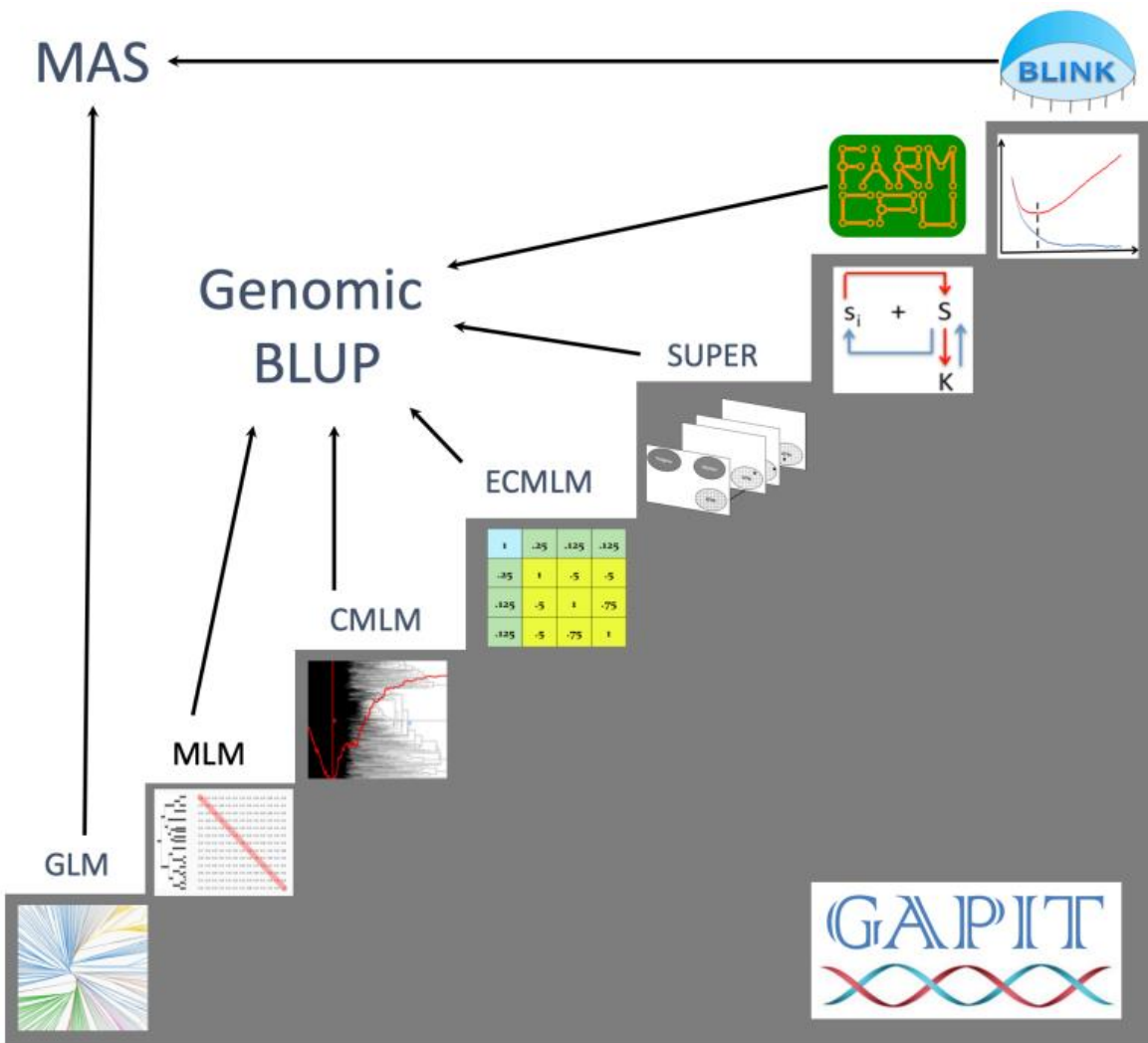


Note: the order of the markers ('probe') in myGM file must be identical with the order of the columns in the myGD file.

The basic GAPIT call:



Which **model** to choose?



Selection criteria:

Computing **efficiency** and statistical **power**?

The various methods sorted according to their statistical power. GLM providing the lowest and BLINK the highest.

The various models are (to some extent) described in the manual.

Proper comparisons between the models is lacking.

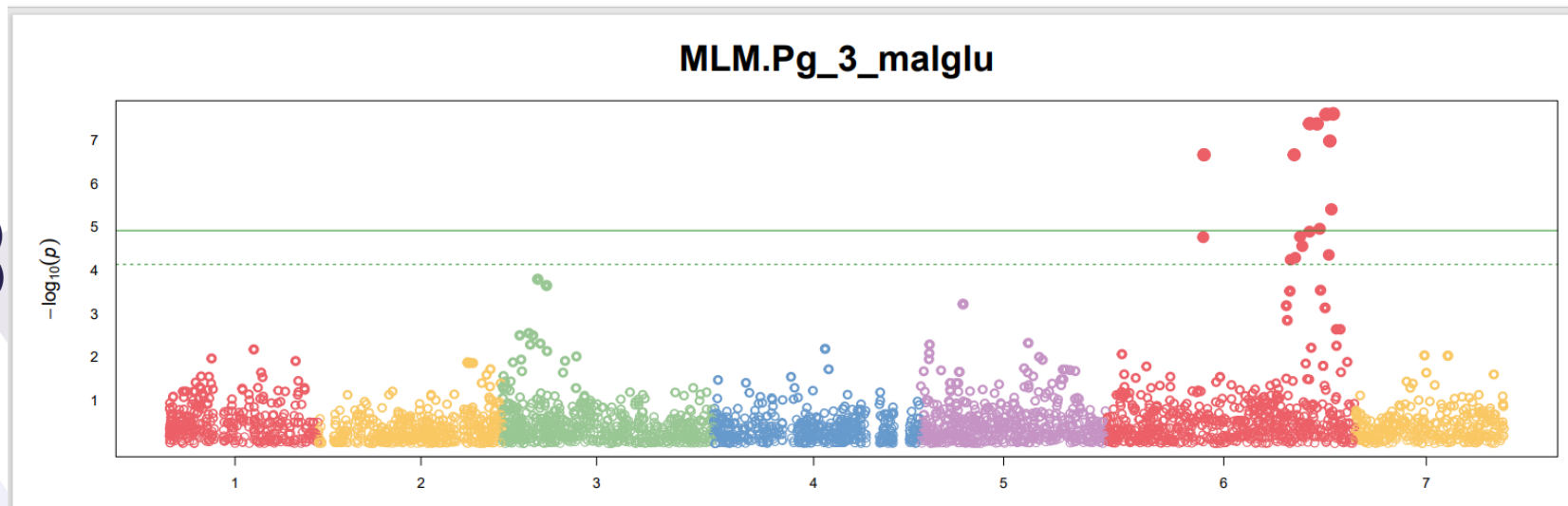
Table 1 Characteristics of methods in GAPIT3

Method	Testing marker	No. of steps	Model	Kinship
GLM	Single locus	One	Fixed	NA
MLM	Single locus	One	Mixed	All markers
CMLM	Single locus	One	Mixed	Individuals clustered into groups
ECMLM	Single locus	One	Mixed	Individuals clustered into groups by enrichment
SUPER	Single locus	Two	Mixed	All marker except pseudo QTNs
MLMM	Multiple loci	Iterative	Mixed	All markers
FarmCPU	Multiple loci	Iterative	Fixed and mixed	Pseudo QTNs
BLINK	Multiple loci	Iterative	Fixed	NA
gBLUP	NA	One	Mixed	All markers for all individuals
cBLUP	NA	One	Mixed	Individuals clustered into groups with all markers
sBLUP	NA	One	Mixed	Pseudo QTNs

Note: NA, not applicable; GLM, general linear model; MLM, mixed linear model; CMLM, compressed MLM; ECMLM, enrichment CMLM; SUPER, settlement of MLMs under progressively exclusive relationship; MLMM, multiple loci MLM; FarmCPU, fixed and random model circulating probability unification; BLINK, Bayesian-information and linkage-disequilibrium iteratively nested keyway; gBLUP, genomic best linear unbiased prediction; cBLUP, compressed BLUP; sBLUP, SUPER BLUP; QTN, quantitative trait nucleotide.

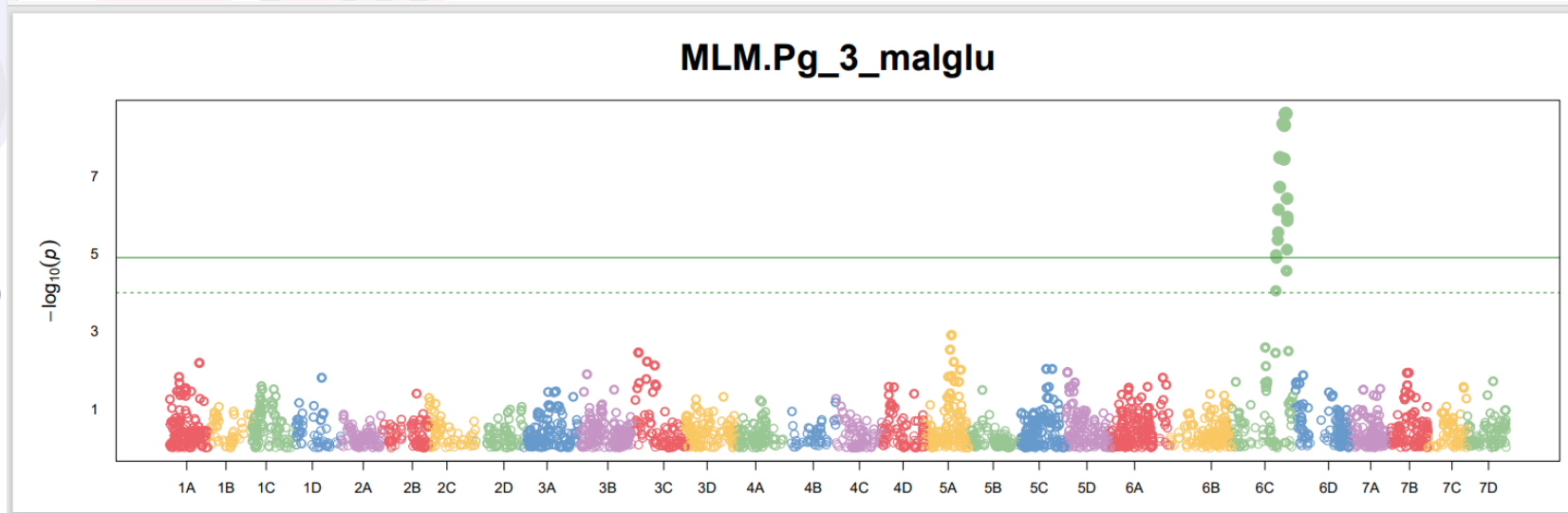
ManhattanPlot with the F.vesca H4.0 genome and iStraw35 markers

Bonferroni (0.01)
Bonferroni (0.05)

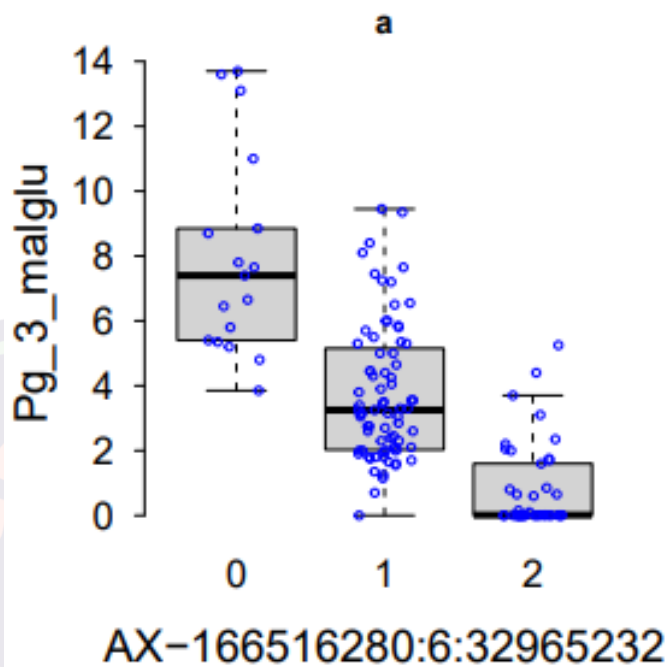


ManhattanPlot with the Royal Royce genome and iStraw35 markers.

Bonferroni (0.01)
Bonferroni (0.05)

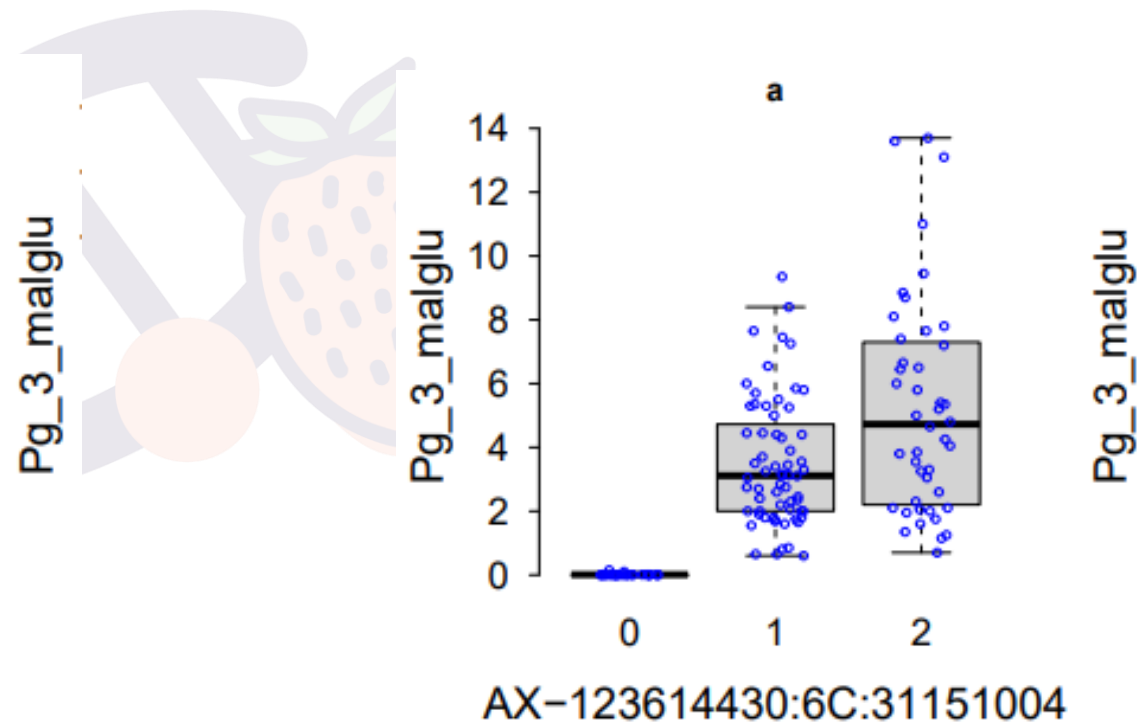


With the *F.vesca* H4.0 genome and iStraw35 markers.



Most significant (BLINK):
AX-166516280 at 32.97 MB

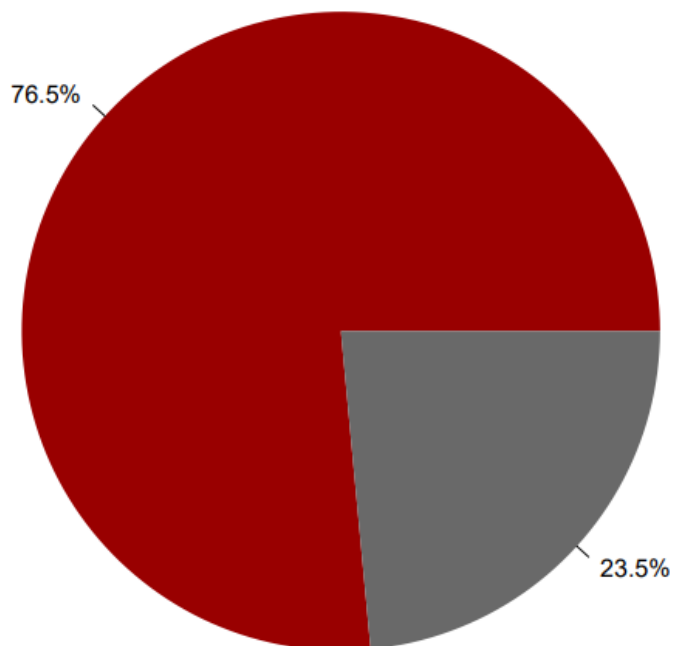
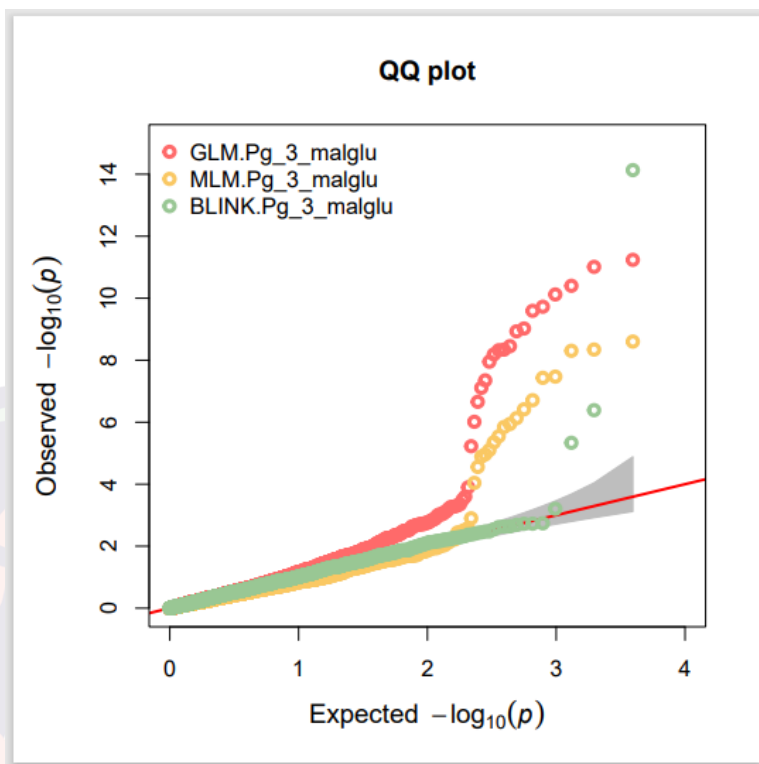
With the Royal Royce genome and iStraw35 markers.



Most significant (BLINK):
AX-123614430 at 31.15 MB

$$R^2 = 2 \times \text{MAF} \times (1 - \text{MAF}) \times \beta^2 = 2 \times 0.43 \times (1 - 0.43) \times 2.89^2 = 0.88$$

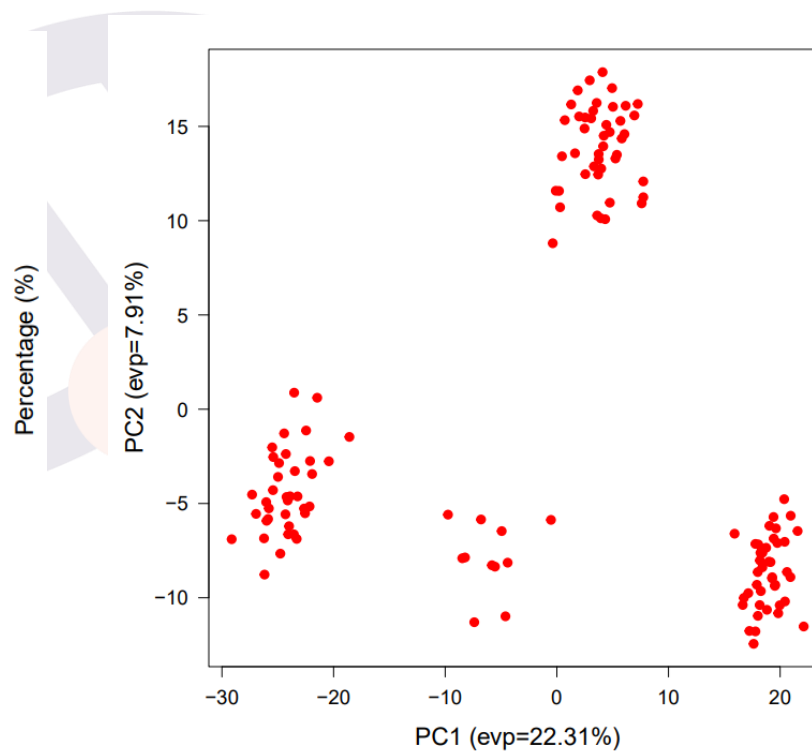
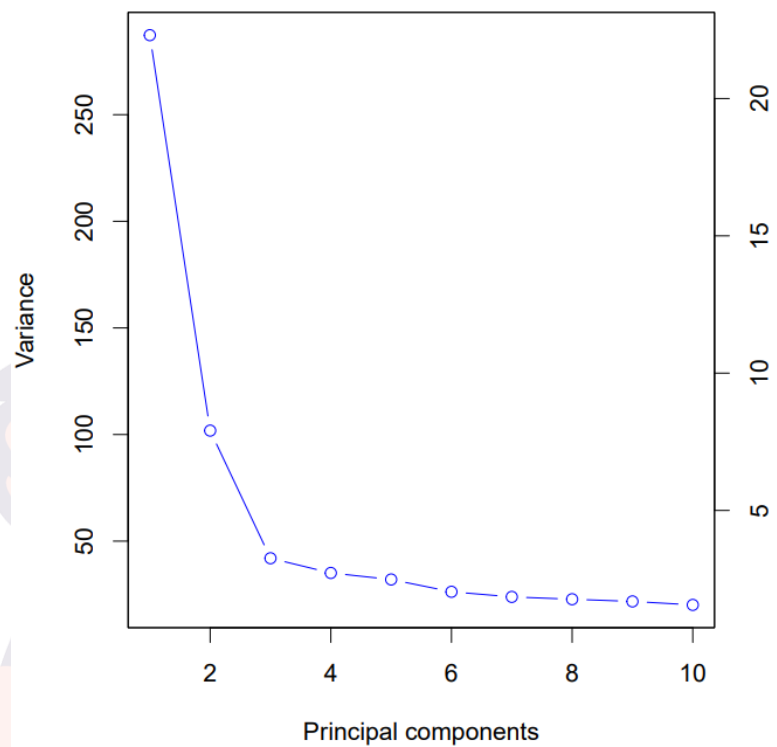
Percentage of variation in pelargonidin-3-*O*-malonylglucoside explained by markers.



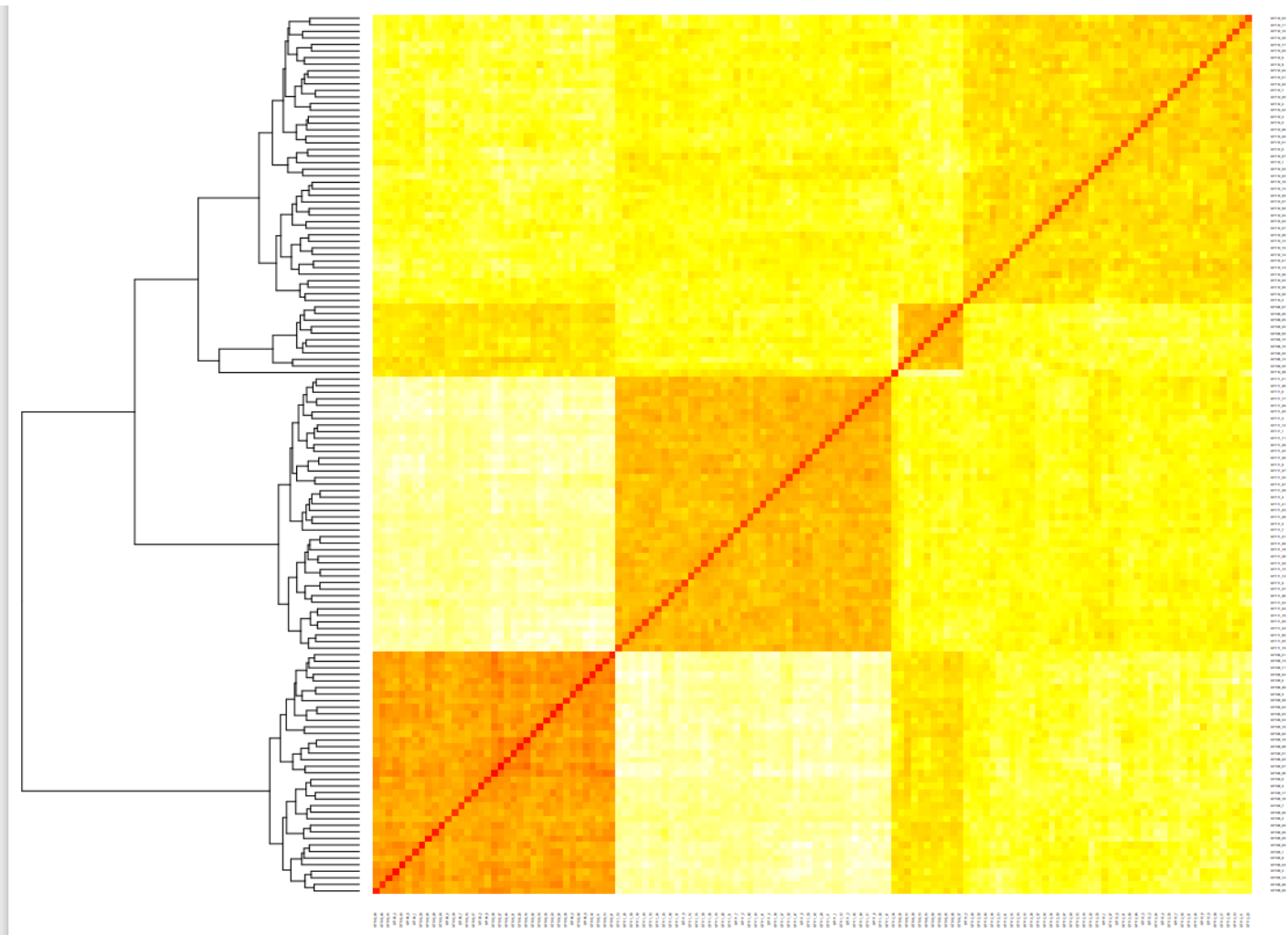
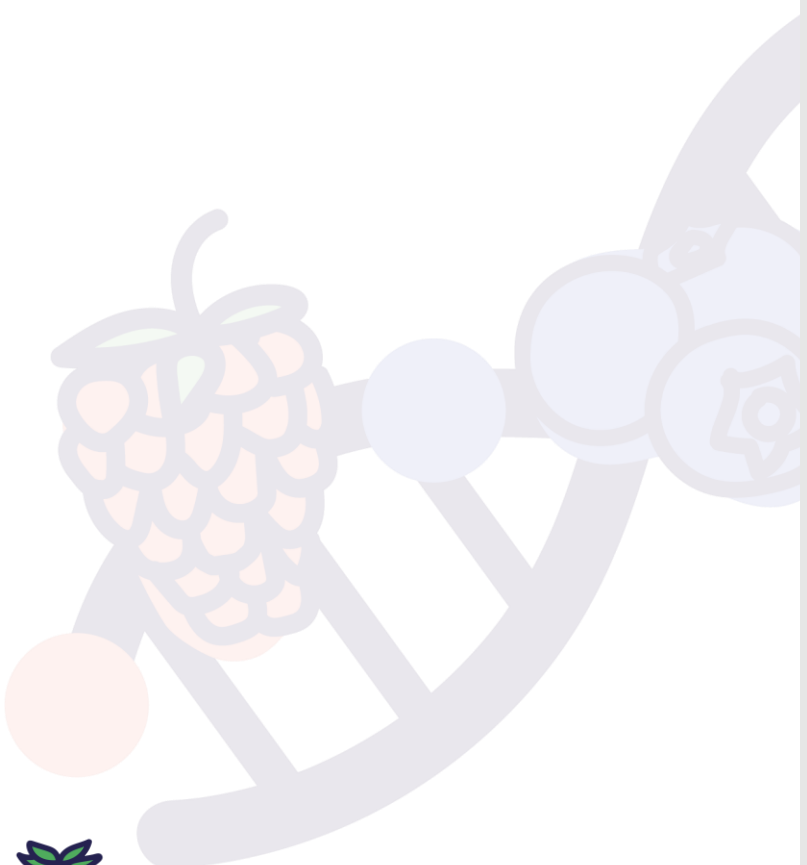
■ Genetic: 7.29
■ Residual: 2.24

The optimum compression
Cluster method: Mean
Group method: average
Group number: 134
-2LL: 598.38

How many principal components to include?



Contaminants?
most deff



What is it good for?

- Homing in on functional genes.
- Marker assisted selection
- Use the markers as covariates in GS

BTW:

- GAPIT has several GS tools implemented
 - gBLUP → large number of genes
 - sBLUP → small number of genes
 - cBLUP → for traits with low H^2

The current data set (the iStraw35 markers and the phenotypes) and R-code is available upon request (jahn.davik@nibio.no)
The GAPIT User Manual is very instructive, and a tutorial data set is also available there.

Thank you for the attention.