Lecture 3 - Basics of GWAS and signal detection

Alencar Xavier, Gota Morota

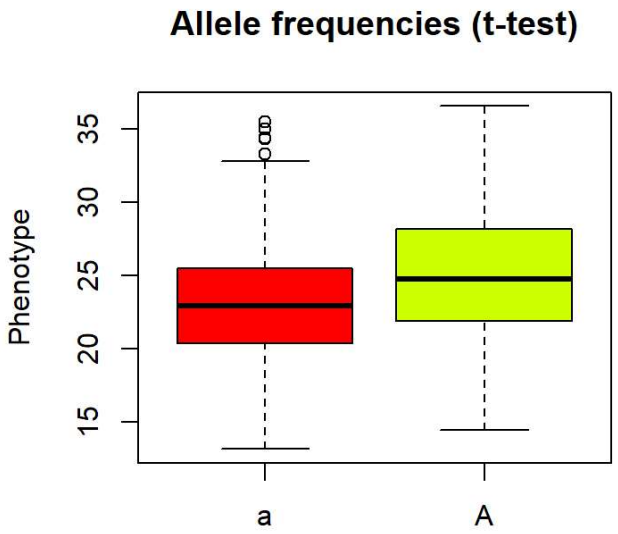
October 26, 2018

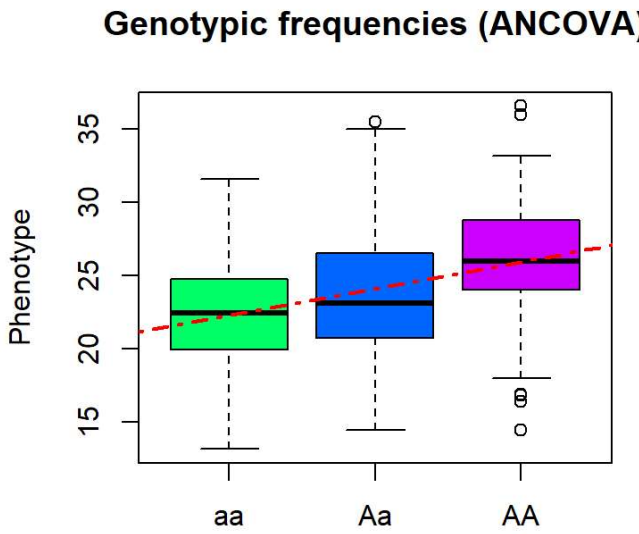
Outline

- Test statistics
- Allele coding
- Power & resolution
- Linkage mapping
- LD mapping
- Structure
- Imputation
- GLM
- MLM
- WGR
- Rare-variants
- Validation studies

Test statistics

- Testing associations are as simple as t-test and ANOVA





Test statistics

- A more generalized framework: Likelihood test

$$LRT = L_0/L_1 = -2(logL_1 - logL_0)$$

For the model:

$$y = Xb + Zu + e$$
$$y \sim N(Xb, V)$$

REML function is given by:

$$L(\sigma_u^2, \sigma_e^2) = -0.5(ln|V| + ln|X'V^{-1}X| + y'Py)$$

Where $V = ZKZ'\sigma_u^2 + I\sigma_e^2$ and $y'Py = y'e$

Allele coding

Types of allele coding

1. Add. (1 df): {-1,0,1} or {0,1,2} - **Very popular** (Lines, GCA)
2. Dom. (1 df): {0,1,0} - **Popular** (Trees, clonals and Hybrids)
3. Jointly A+D (2 df): Popular on QTL mapping in F2s
4. Complete dominance (1 df): {0,0,1} or {0,1,1} - **Very unusual**
5. Interactions (X df): (epistasis and GxE)

Power and resolution

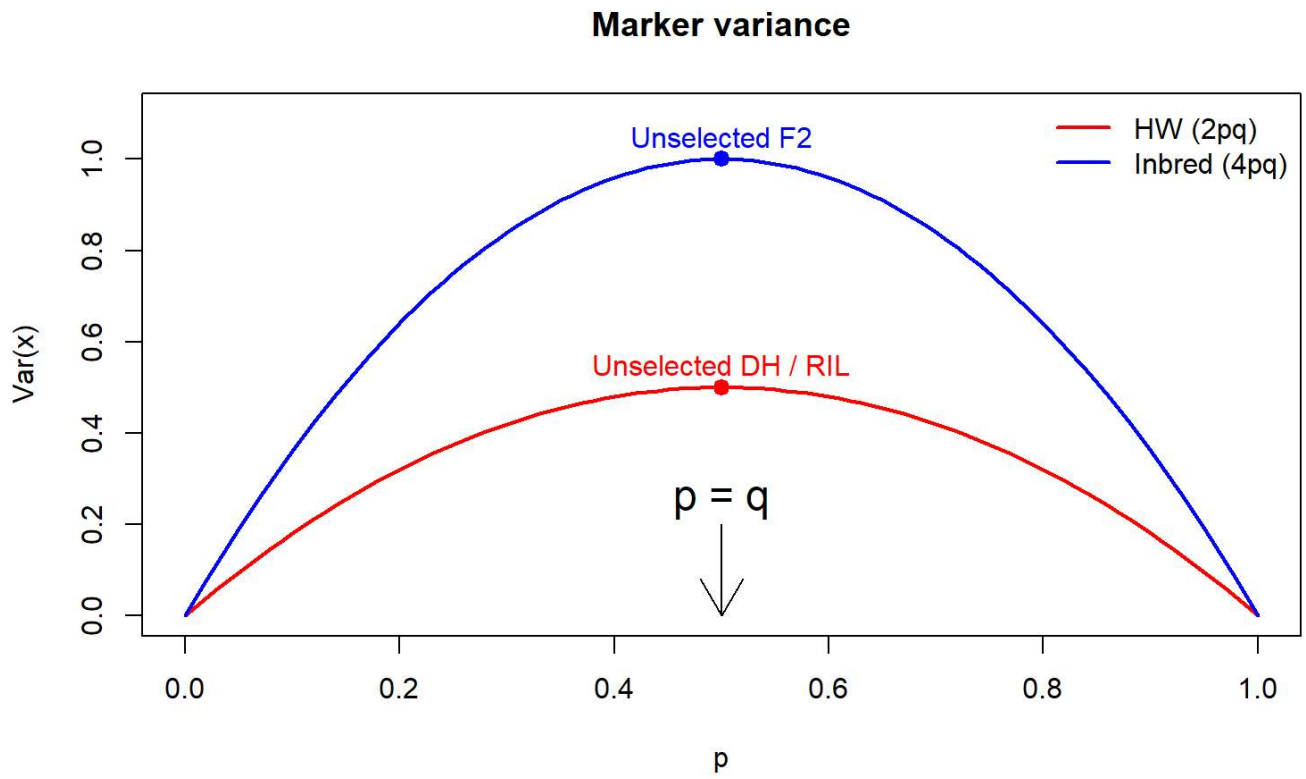
Power

- Key: Number of individuals & allele frequency
- More DF = lower power
- Multiple testing: Bonferroni and FDR
- Tradeoff: Power vs false positives

Resolution

- Genotyping density
- LD blocks
- Recombination

Power: Variance of X



Beavis effect: 1000 is just OK

Xu S. Theoretical basis of the Beavis effect. Genetics. 2003 1;165(4):2259-68.

TABLE 4

Comparisons of predicted and observed (estimated) biases in estimated QTL effects and variances from Beavis F_2 simulation experiments

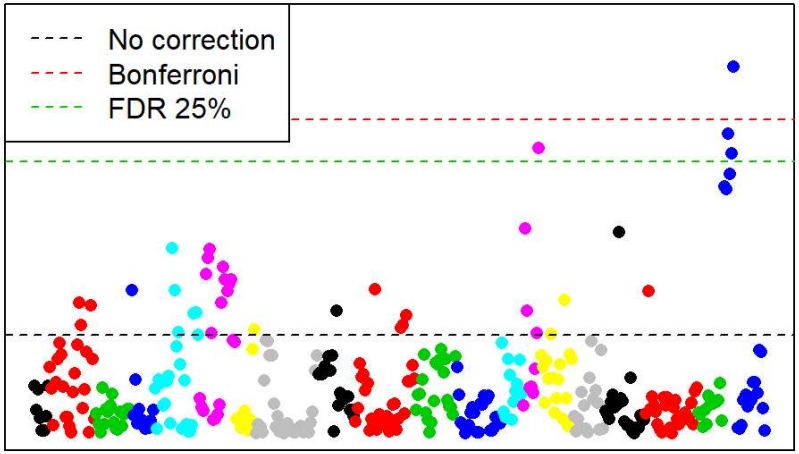
Simulated conditions[a]	Variance explained			Additive effect			Average estimated location
	Simulated	Observed	Predicted[b]	Simulated	Observed	Predicted[c]	
10-30-100	3.00	16.76	16.0537	2.45	4.96	5.6410	11.3
10-30-500	3.00	4.33	4.1890	2.45	2.89	2.8617	10.53
10-30-1000	3.00	3.02	3.1846	2.45	2.56	2.4868	10.8
10-63-100	6.25	12.65	16.5984	3.55	4.68	5.7328	10.51
10-63-500	6.25	7.08	6.5581	3.55	3.73	3.5829	10.96
10-63-1000	6.25	6.34	6.3566	3.55	3.60	3.5500	11.04
10-95-100	9.50	18.68	17.3883	4.36	5.85	5.8466	10.58
10-95-500	9.50	10.1	9.7082	4.36	4.49	4.3607	11.08
10-95-1000	9.50	9.67	9.6028	4.36	4.44	4.3600	11.19
40-30-100	0.75	15.78	15.6270	1.22	4.40	5.5436	10.83
40-30-500	0.75	3.17	3.3332	1.22	2.35	2.5671	10.17
40-30-1000	0.75	1.46	1.7961	1.22	1.85	1.8790	10.17
40-63-100	1.56	16.31	15.7983	1.77	4.71	5.5999	10.45
40-63-500	1.56	3.54	3.5783	1.77	2.59	2.6582	10.13
40-63-1000	1.56	1.96	2.1435	1.77	2.09	2.0494	10.37
40-95-100	2.40	16.55	15.9694	2.18	5.02	5.6236	10.45
40-95-500	2.40	3.97	3.9190	2.18	2.79	2.7641	10.12
40-95-1000	2.40	2.58	2.6970	2.18	2.36	2.2784	10.29

[a] Numerical values denote the number of QTL-heritability-number of progeny.
[b] Using Equation 17.
[c] Using Equation 8.

Multiple testing:

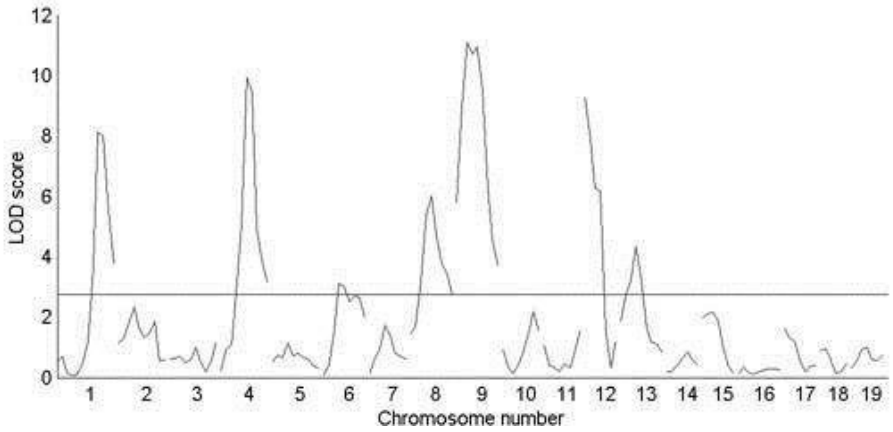
GWAS tests m hypothesis:

- No correction: $\alpha = 0.05/m$
- Bonferroni: $\alpha = 0.05/m$
- FDR (25%): $\alpha = 0.05/(m \times 0.75)$



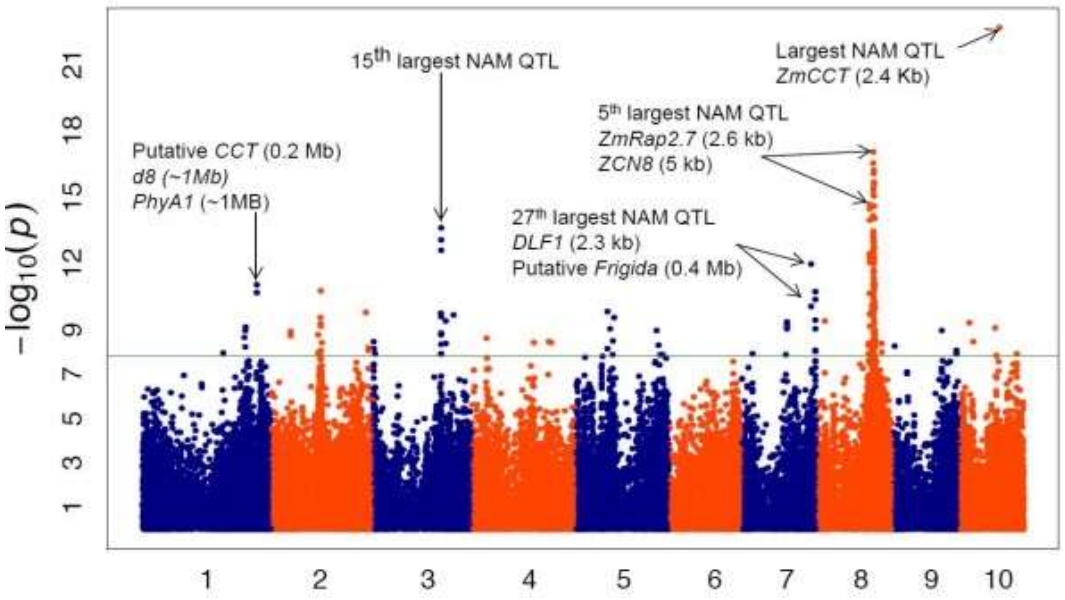
Linkage mapping

- Generally on experimental pops (F2, DH, RIL, BC)
- Based on single-marker analysis or interval mapping
- Recombination rates would increase **power**



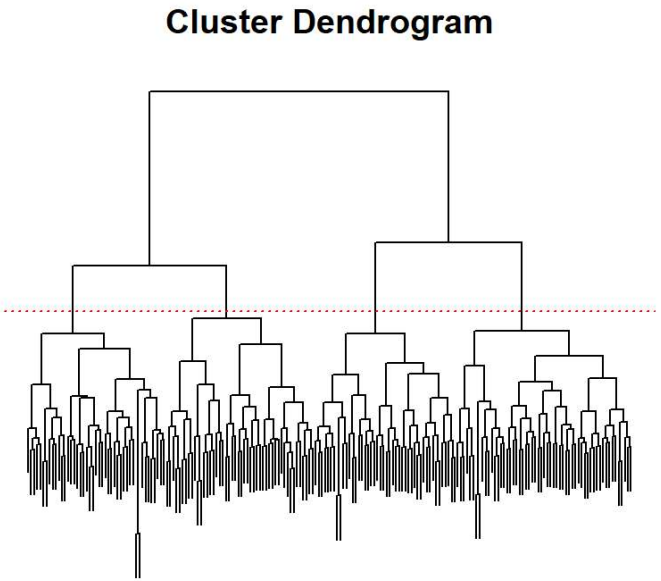
LD mapping (or association mapping)

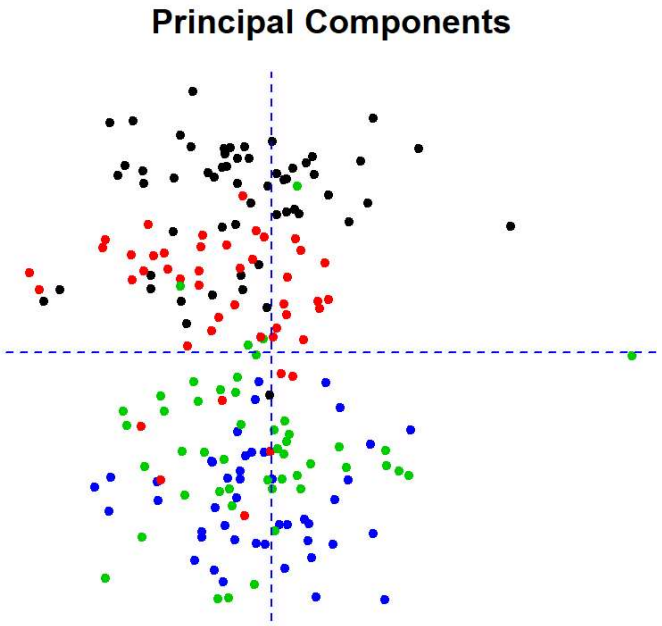
- Use of historical LD between marker and QTL
- AM allowed studies on random panels
- Dense SNP panels would increase **resolution**



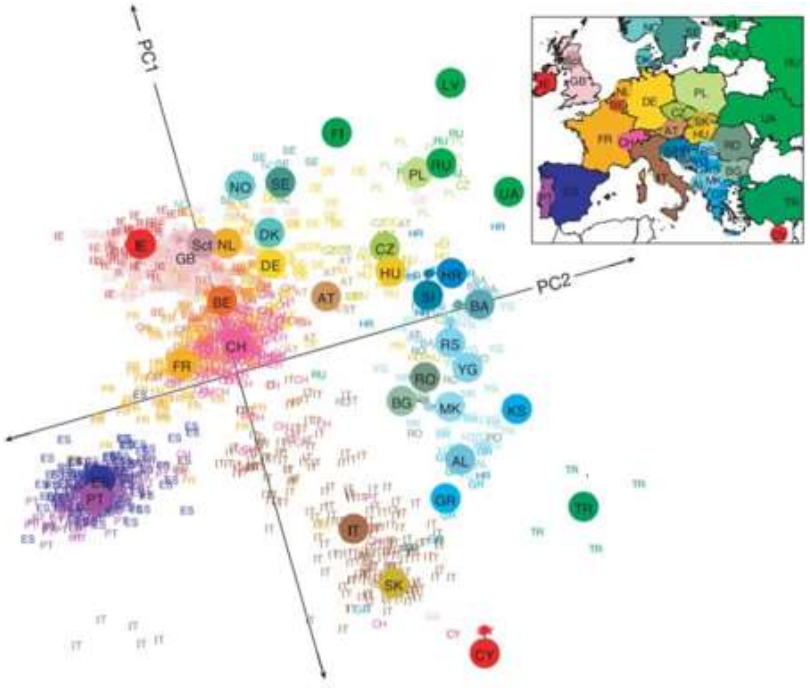
Structure

1. Confounding associations with sub-populations
2. Major limitation of association mapping
3. Structure: , , (eg. race)





Structure



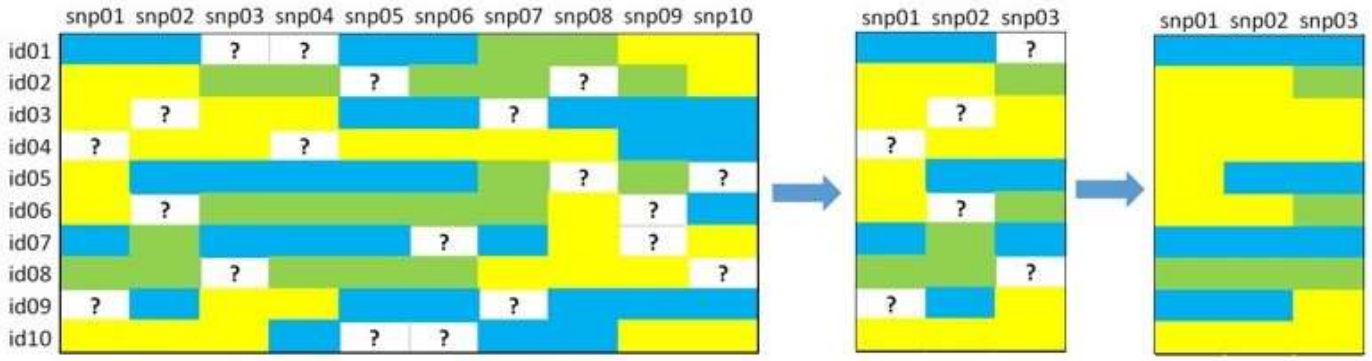
Genes mirror geography within Europe

nature Vol 456|6 November 2008|doi:10.1038/nature07331

Imputation

Less missing values = more obs. = more detection power

- Markov models: Based on flanking markers
- Random forest: Multiple decision trees capture LD
- kNN & Projections: Fill with similar haplotypes



GLM (generalized linear models)

- Full model (L_1):

$$y = Xb + m_j a + e$$

- Null model (L_0):

$$y = Xb + e$$

1. **Advantage**: Fast, not restricted to Gaussian traits
2. Popular methodology on human genetic studies
3. Xb includes (1) environment, (2) structure and (3) covariates

MLM (mixed linear models)

- Full model (L_1):

$$y = Xb + Zu + m_j a + e$$

- Null model (L_0):

$$y = Xb + Zu + e$$

1. The famous "**Q+K model**"
2. **Advantage**: Better control of false positives, no need for PCs
3. Polygenic effect (u) assumes $u \sim N(0, K\sigma_u^2)$
4. Faster if we don't reestimate $\lambda = \sigma_e^2/\sigma_u^2$ for each SNP

cMLM (compressed MLM)

1. Uses the same base model as MLM
2. **Advantage**: Faster than MLM
3. Based on clustered individuals:

- Z is indicates the subgroup
- K is the relationship among subgroup
- Often needs PCs to complement K

WGR (whole-genome regression)

1. Tests all markers at once
2. **Advantage**: No double-fitting, no PCs, no Bonferroni

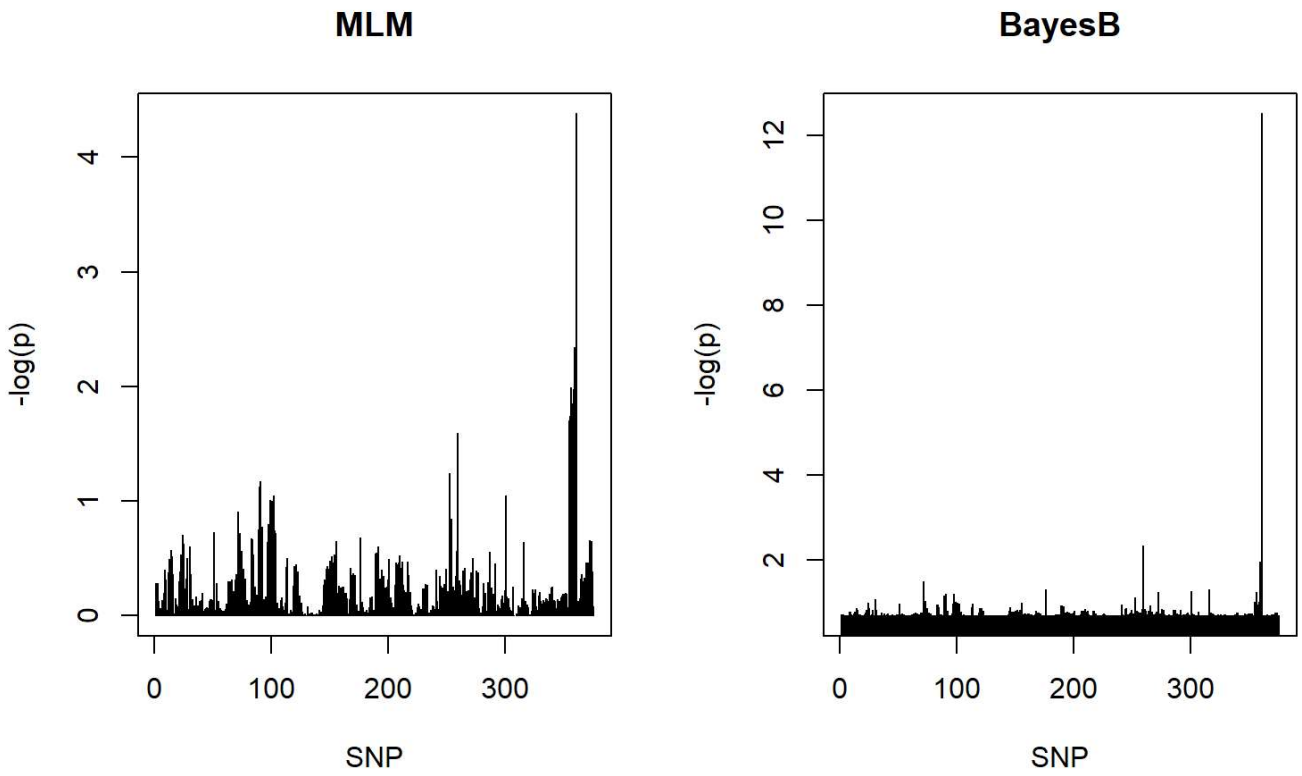
- Model (BayesB, BayesC, SSVS):

$$y = Xb + Ma + e$$

- Marker effects are from a mixture of distributions

$a_j \sim Binomial$ with $p(\pi) = 0$ and $p(1 - \pi) = a_j$

WGR (whole-genome regression)



Rare variants

1. Screen a set (s) of low MAF markers on NGS data
2. **Advantage**: Detect signals from low power SNPs
3. Applied to uncommon diseases (not seen in plant breeding)
4. Two possible model

- Full model 1 (L_1): $y = Xb + M_s a + e$
- Full model 2 (L_2): $y = Xb + PC_1(M_s) + e$
- Null model (L_0): $y = Xb + e$

Test either $LR(L_1, L_0)$ or $LR(L_2, L_0)$

Validation studies

- QTLs detected with 3 methods, across 3 mapping pops
- Validations made on 3 unrelated populations

https://doi.org/10.1534/g3.118.200636

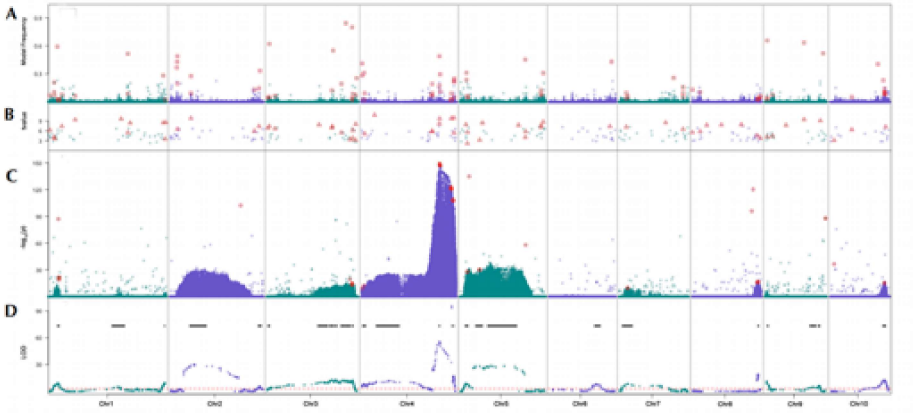


Figure 2 Stacked plots of GWAS and QTL results. From upper to lower panels are results from the Bayesian-based multi-variant (A) stepwise regression (B) and single variant(C) models for GWAS and the joint QTL mapping result (D). The red dashed line in the QTL plot indicates the 1,000 permutation threshold and black lines show the QTL confidence intervals. Red squares in panel (A), triangles in panel (B) and circles in panel (C) indicate the kernel row number associated variants selected for further genetic validation.