

Defining the causes and consequences of phantom epistasis in heterotic hybrids

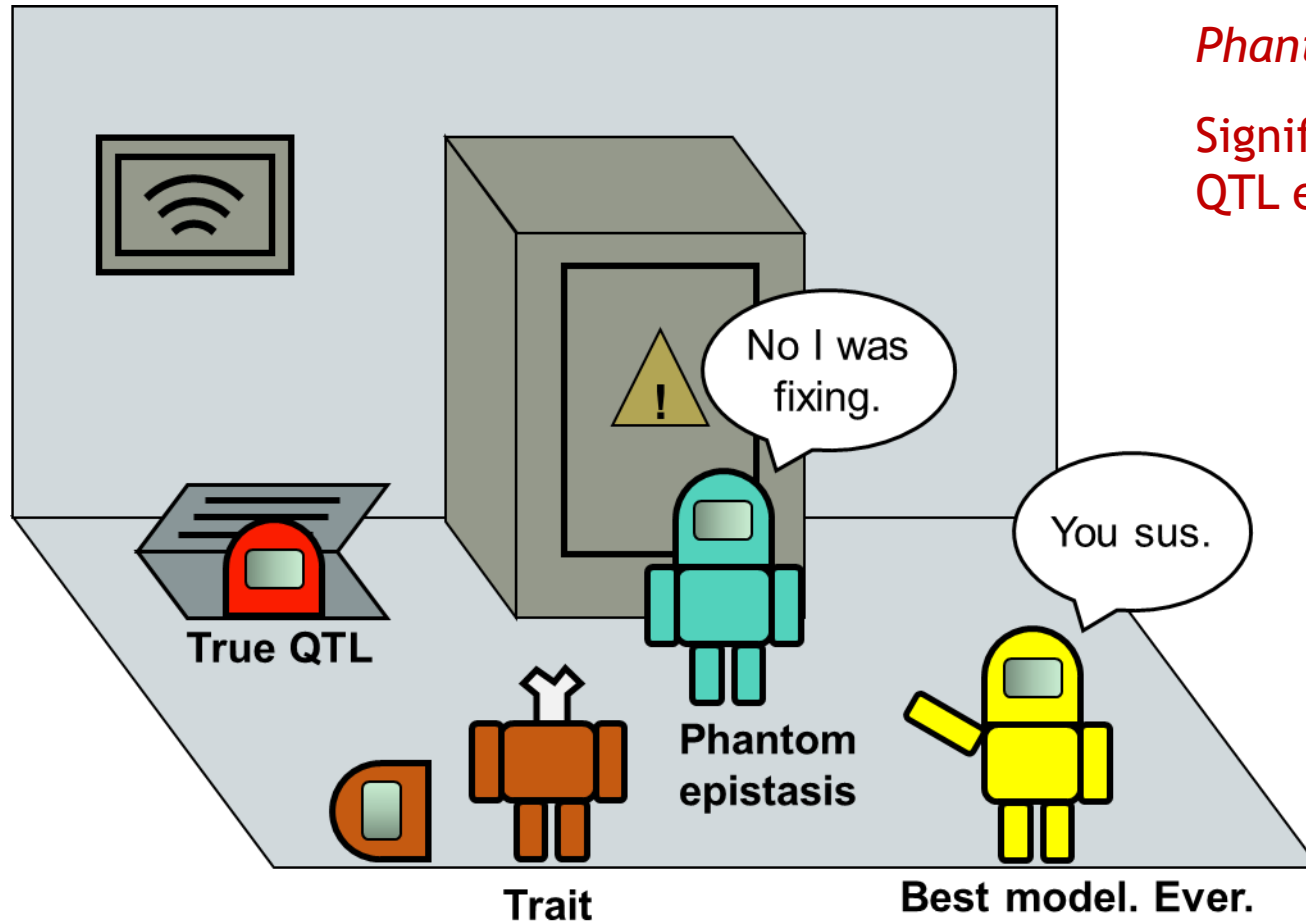
CJ Yang, Wayne Powell, Ian Mackay

Computational Genomics Discussion Group (CGDG) Seminar

22nd November 2022

1. Definition and example of phantom epistasis.
2. Quick introduction to heterosis.
3. Simulated distribution of mid-parent heterosis.
4. Simulating traits of various genetic architecture.
5. Estimating variance components.
6. Single locus and interaction GWAS.

Reboot version of my previous talk at the 18th Eucarpia Biometrics in Plant Breeding Conference.



Phantom/apparent epistasis:

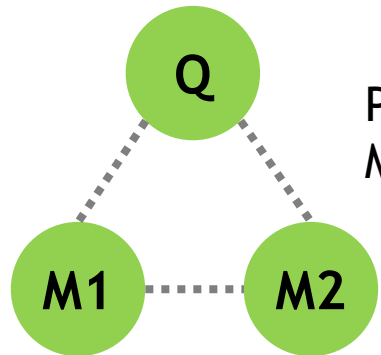
Significant epistatic effect when the causative QTL effect is purely additive and/or dominance.

One possible cause of phantom epistasis

Imperfect linkage disequilibrium generates phantom epistasis (& perils of big data).

de los Campos et al (2019)

<https://doi.org/10.1534/g3.119.400101>



Phantom epistasis can arise between M1 and M2 if there are imperfect LDs.

Example of phantom epistasis

Detection and replication of epistasis influencing transcription in humans.

Hemani et al (2014)

<https://doi.org/10.1038/nature13005>



Another explanation for apparent epistasis.

Wood et al (2014)

<https://doi.org/10.1038/nature13691>



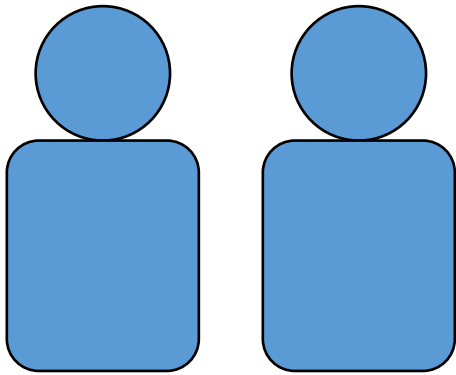
Phantom epistasis between unlinked loci.

Hemani et al (2021)

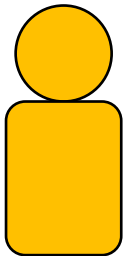
<https://doi.org/10.1038/s41586-021-03765-z>



Heterosis: F1 progeny outperforms the parents.



Kiddo, you will live a better life than us when you grow up.



1. Best Parent Heterosis (BPH) = $F1 > P1, P2$.
2. Mid Parent Heterosis (MPH) = $F1 > (P1 + P2)/2$.
3. ~~Worst Parent Heterosis (WPH) = $P1 > F1 > P2$.~~

Heterosis

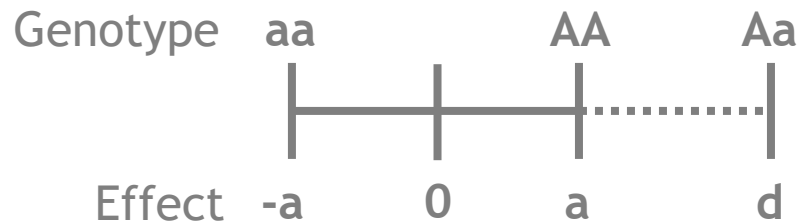


Inbreeding
Depression



Often thought as two sides of the same coin.

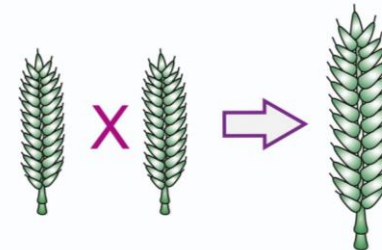
Overdominance



Dispersed dominance (pseudo-overdominance).

(e) Heterosis is also explained by dispersed dominant alleles

genotype: $AAbb \times aaBB$ \rightarrow $AaBb$
phenotype: $1 \times 1 \rightarrow 1.5$



$AA=BB=1$ $aa=bb=0$ $Aa=Bb=0.75$

Mackay et al (2020) CC BY 4.0

Understanding the classics: the unifying concepts of transgressive segregation, inbreeding depression and heterosis and their central relevance for crop breeding.

Mackay et al (2020)

<https://doi.org/10.1111/pbi.13481>



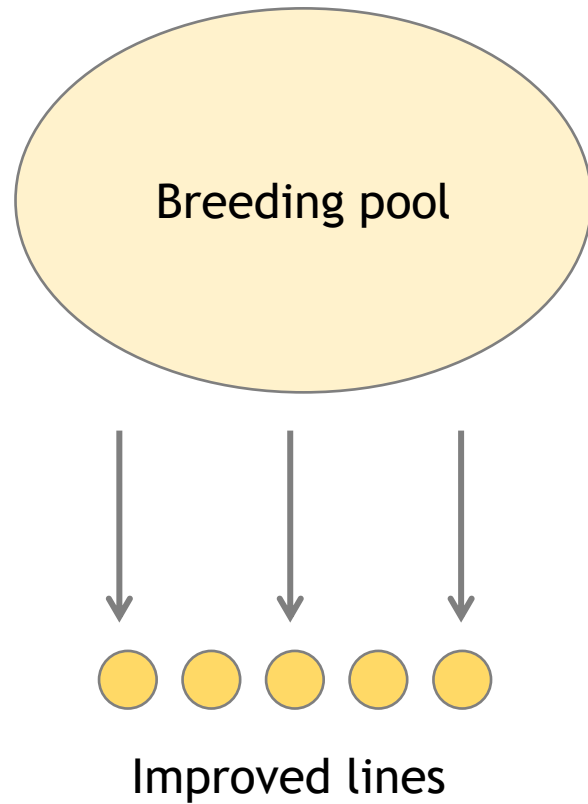
Table 1 Examples of proposed functional mechanisms for heterosis

Mechanism or predictor	Reference
Mitochondrial complementation	Sarkissian and Srivastava (1967)
Metabolic balance	Hageman <i>et al.</i> (1967)
Chloroplast complementation	Srivastava (1981)
Phytohormones, gibberellic acid	Rood <i>et al.</i> (1988)
DNA methylation	Tsaftaris (1995)
Association transcriptomics	Stokes <i>et al.</i> (2010)
Cryptic variation in gene expression	Rosas <i>et al.</i> (2010)
Energy-use efficiency, cell cycle time	Goff (2011)
siRNA	Shivaprasad <i>et al.</i> (2012)
sRNA	Barber <i>et al.</i> (2012)
Florigen pathway	Jiang <i>et al.</i> (2013)
Circadian clock-mediated stress responses	Miller <i>et al.</i> (2015)
Circadian clock gene expression	Shen <i>et al.</i> (2015)

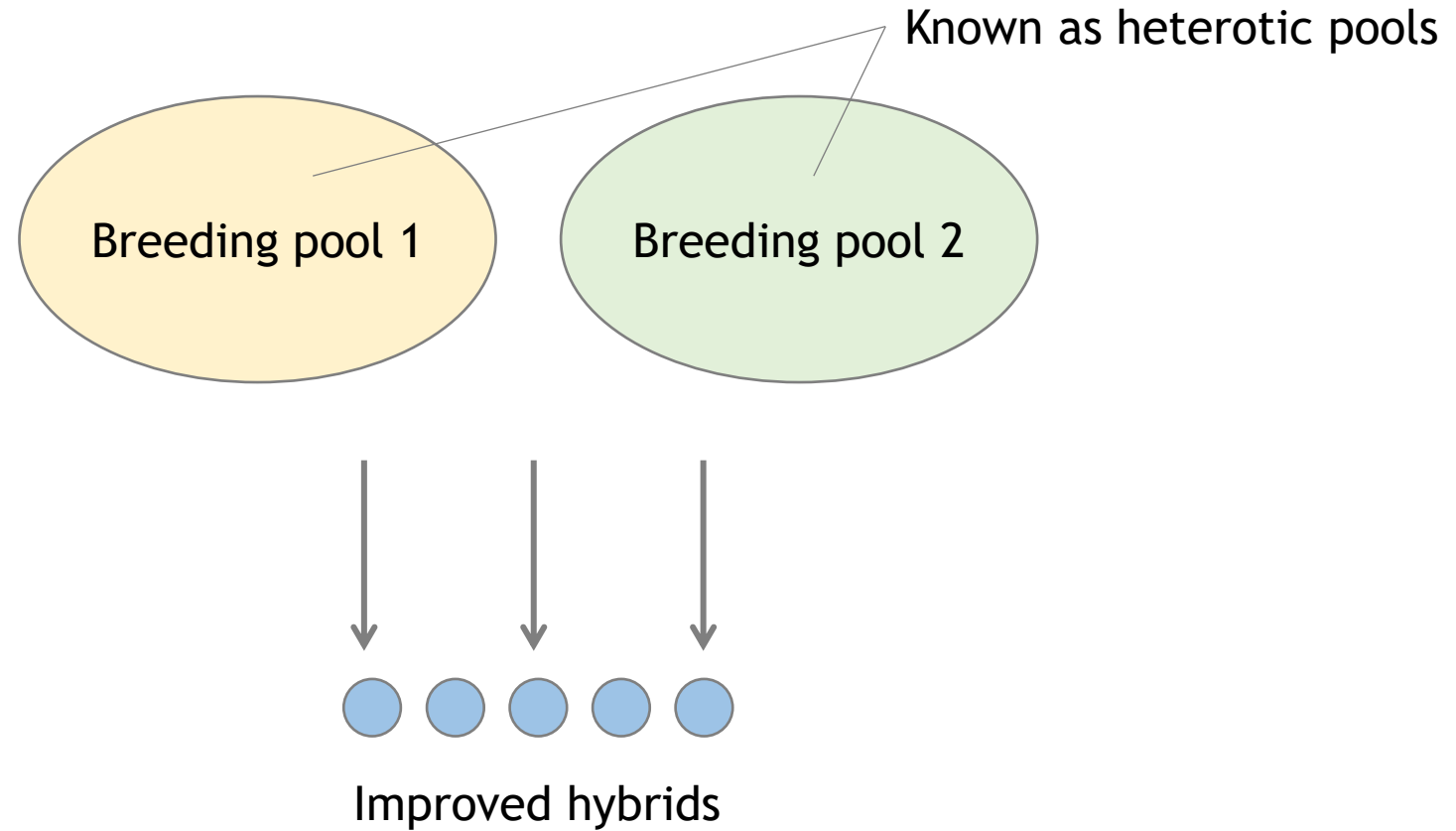
Often seen as isolated cases.

How to make use of heterosis?

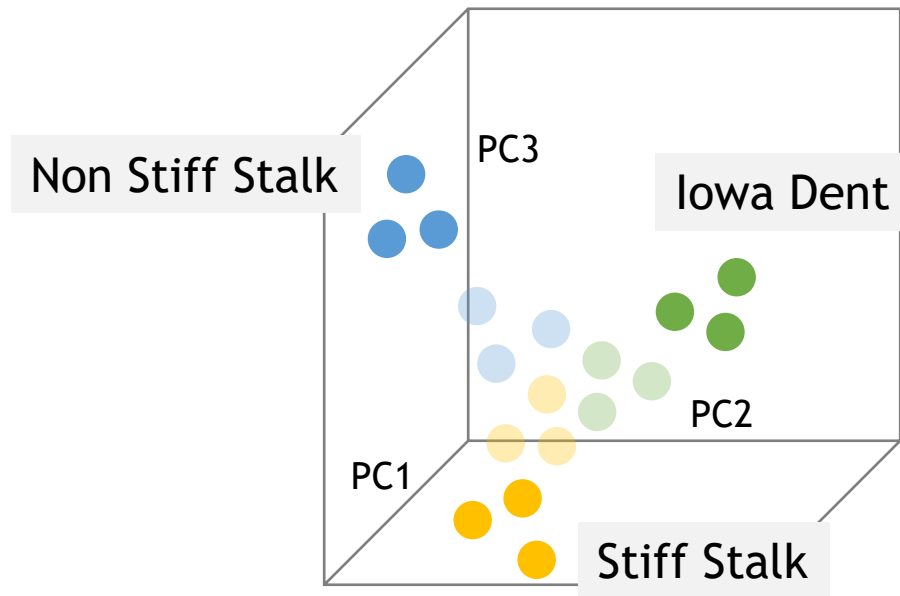
Standard breeding



Hybrid breeding



Heterotic pools have been long established in maize.



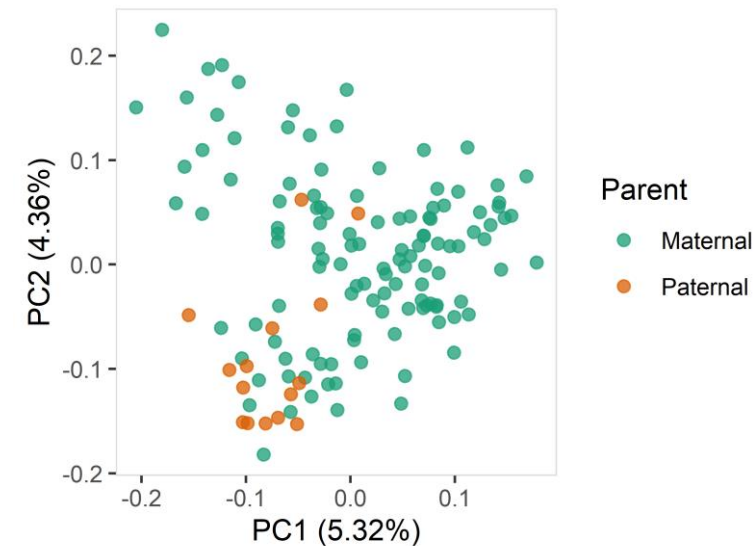
Historical genomics of North American maize.

van Heerwaarden et al (2012)

<https://doi.org/10.1073/pnas.1209275109>



Heterotic pools don't quite exist in wheat, yet.



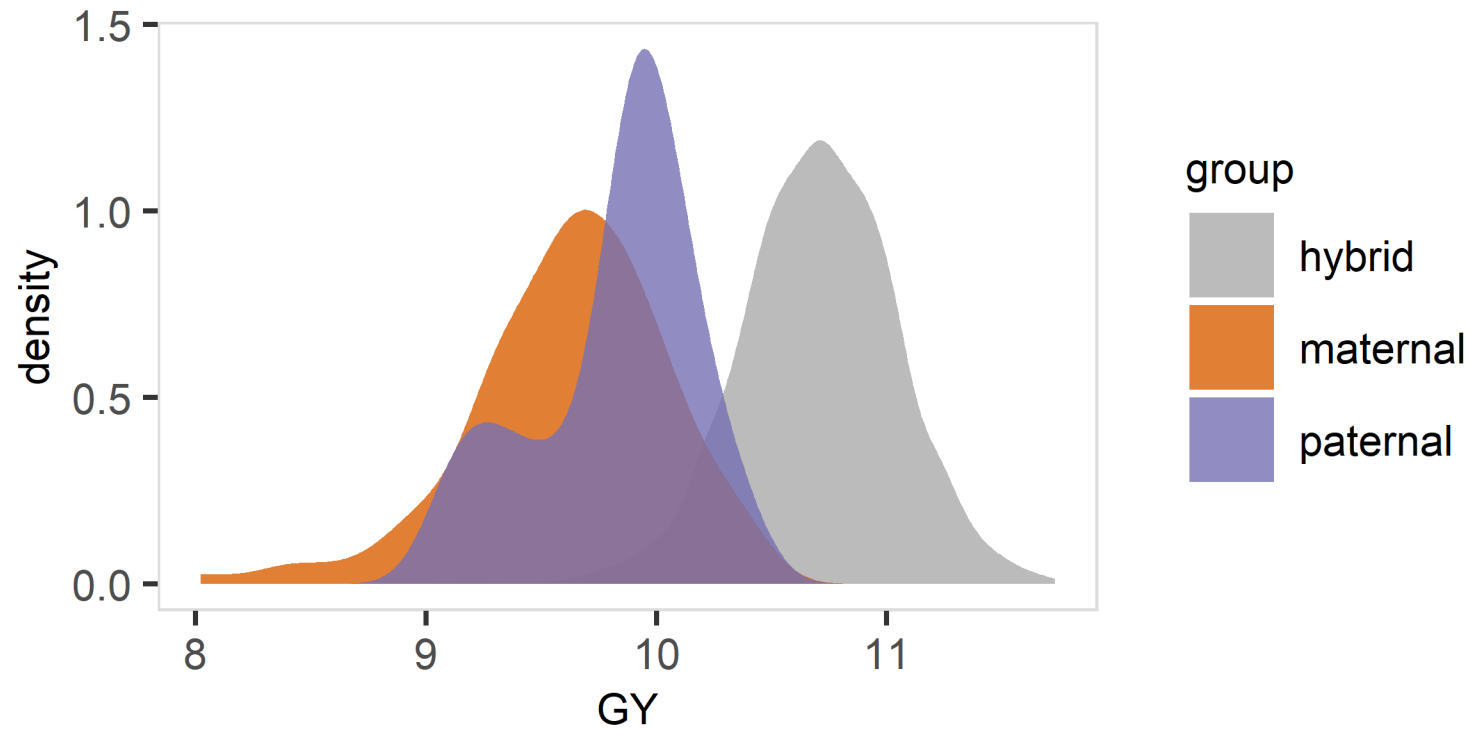
Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding.

Zhao et al (2015)

<https://doi.org/10.1073/pnas.1514547112>

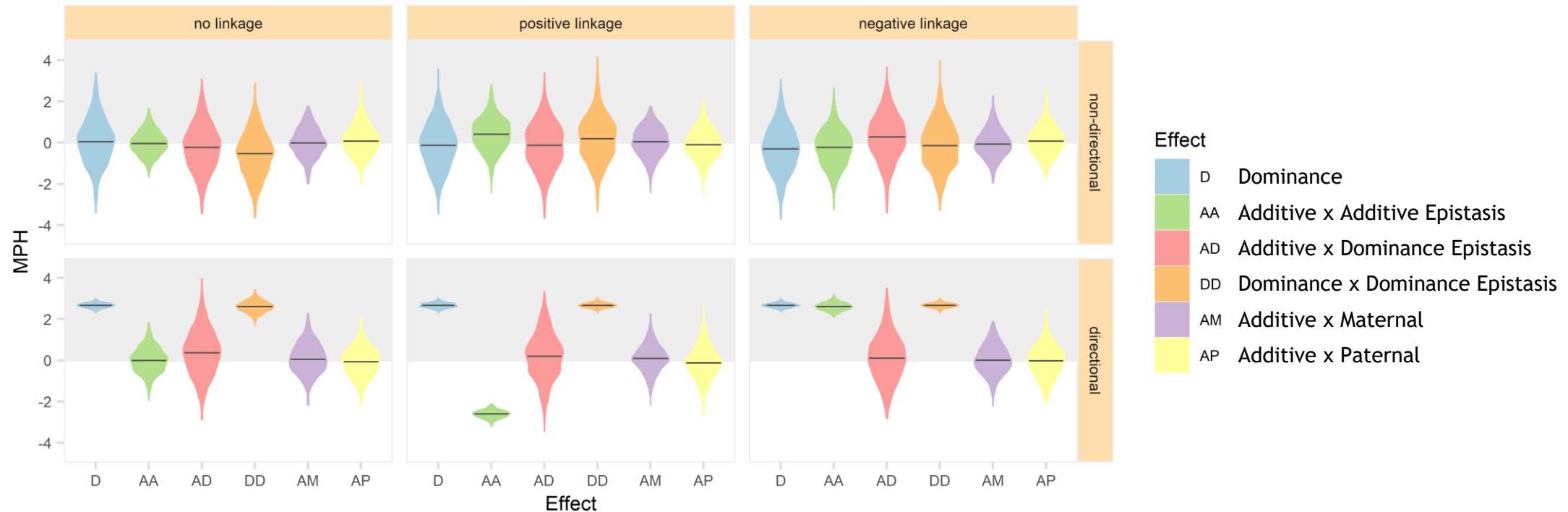



Grain yield (GY) shows ~10% heterosis.



Data from Zhao et al (2015)

Simulated Mid Parent Heterosis (MPH)



- Simulate 6 effects, 1,000 loci ($\times 2$ if linked), 100×2 parents, 1,000 hybrids.
- $MPH = \text{Hybrid} - \text{Parental Average}$.
- Directional effect is needed for non-zero MPH.
- Linkage is needed for non-zero MPH due to AA.  Sounds familiar to “dispersed dominance” ?

A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat.

Jiang et al (2017)

<https://doi.org/10.1038/ng.3974>



Negative dominance and dominance-by-dominance epistatic effects reduce grain-yield heterosis in wide crosses in wheat.

Boeven et al (2020)

<https://doi.org/10.1126/sciadv.aay4897>



Within the genetic variance (V_G) accounting for heterosis:

16/11% Dominance (V_D)

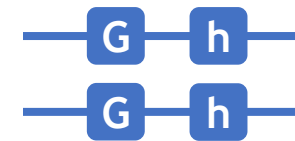
50/61% Additive x additive (V_{AA})

21/17% Additive x dominance (V_{AD})

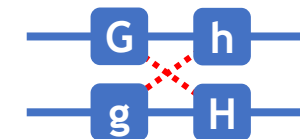
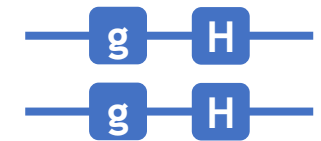
13/11% Dominance x dominance (V_{DD})

Recall that positive MPH from AA requires negative linkage (dispersed favorable AA).

Parent 1 = 0

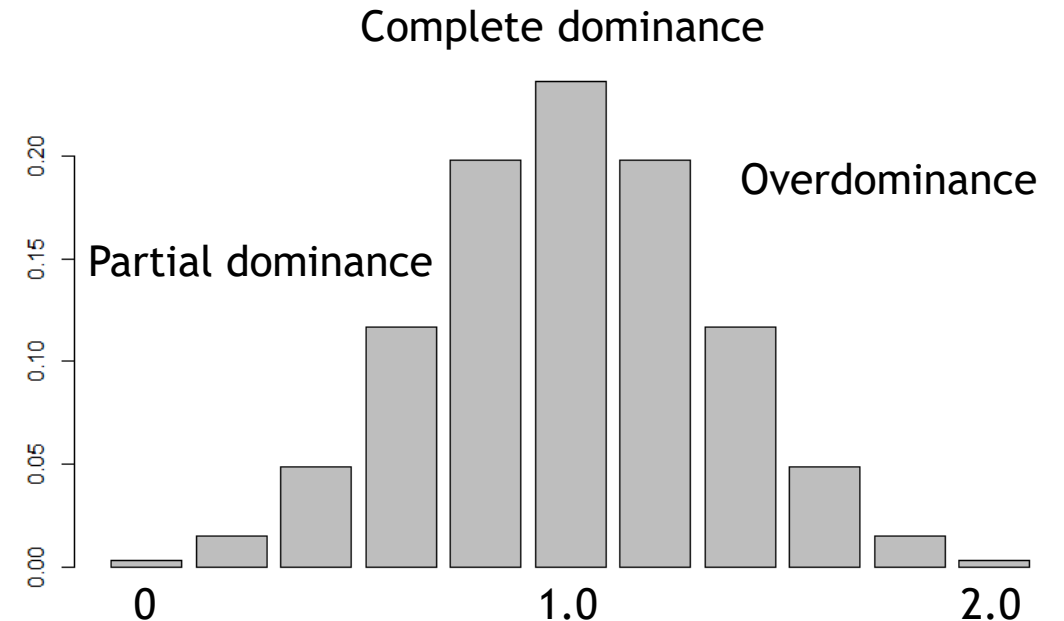


Parent 2 = 0

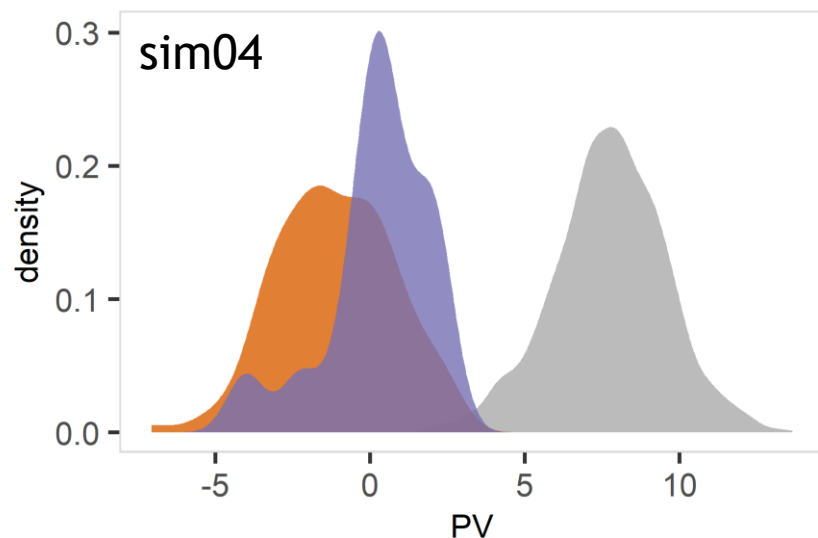


Hybrid = 2

- Hybrid wheat data from Zhao et al (2015).
 - 120 maternal parents + 15 paternal parents.
 - 1,604 hybrids
 - 2,701 markers with positions.
- Simulate 200 pairs of background QTLs with various effects.
- Set $V_{BG,A} = 0.5$ and $V_{res} = 1$.
- Simulate a pair of main QTLs with only A + D.
- Set $V_{Main,A} = 0.5$ and complete dominance.
- All genetic effects are uni-directional.
- Thin markers based on linkage to QTLs.
- 10 simulations for each genetic architecture.

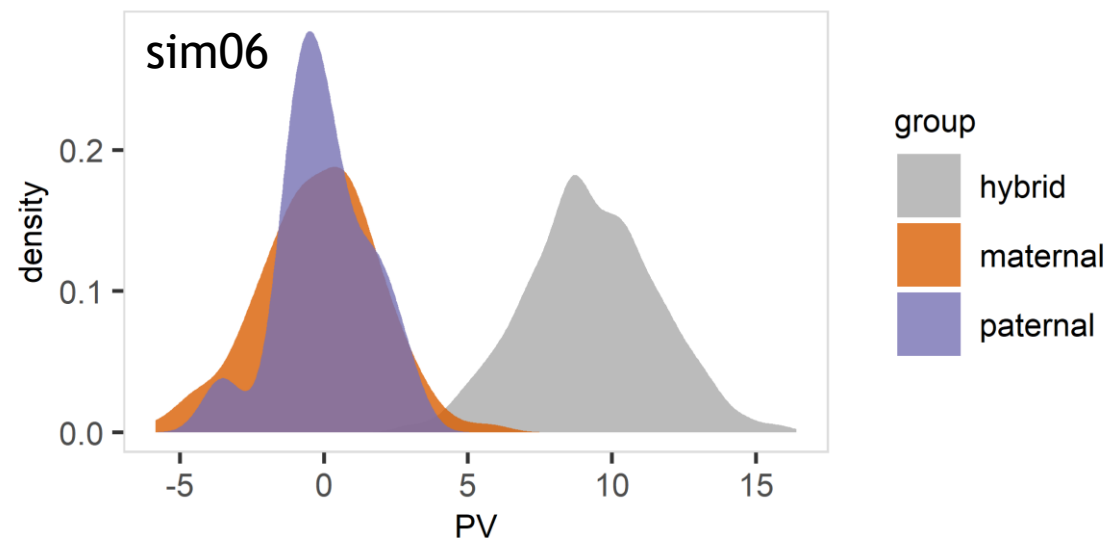


Set other effects by multiplying the additive effects to a random value drawn from this distribution.



A+D

Biological	Statistical
$V_A = 1.20$	$V_A = 1.15$
$V_D = 1.78$	$V_D = 0.45$
$V_{AA} = 0.00$	$V_{AA} = 0.00$
$V_{AD} = 0.00$	$V_{AD} = 0.00$
$V_{DD} = 0.00$	$V_{DD} = 0.00$
$V_{Res} = 1.00$	$V_{Res} = 1.00$



A+D+AA+AD+DD

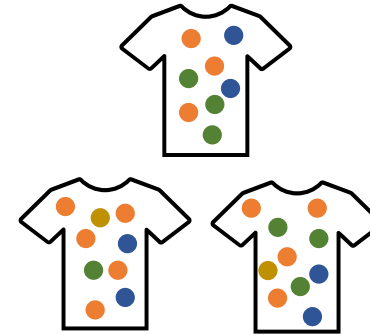
Biological	Statistical
$V_A = 0.85$	$V_A = 2.34$
$V_D = 1.67$	$V_D = 0.57$
$V_{AA} = 0.29$	$V_{AA} = 0.14$
$V_{AD} = 0.29$	$V_{AD} = 0.07$
$V_{DD} = 0.17$	$V_{DD} = 0.01$
$V_{Res} = 1.00$	$V_{Res} = 1.00$

Trait = Mean + Genetic + Residual

$$y = \mu + g_A + g_D + g_{AA} + g_{AD} + g_{DD} + \varepsilon$$

$$g_i \sim N(0, \mathbf{K}_i V_i) \quad i = A, D, AA, AD, DD$$

$$\varepsilon \sim N(0, \mathbf{I} V_{Res})$$



Think of y as a pile of dirty laundry, and we want to figure out the type of dirt (g_i) on each shirt.

And \mathbf{K}_i tells us which clothes have been to the same place.

Model 1: classical mixed model using `mmer()` in R/sommer.

Model 2: Bayesian model using `BGLR()` in R/BGLR.

Both models need \mathbf{K} , i.e. Genomic Relationship Matrices (GRMs).

GRM is typically calculated as:
$$K = \frac{\tilde{M} \cdot \tilde{M}'}{c}$$

Where:

M is a numeric marker genotype matrix (rows as individuals, columns as markers).

\tilde{M} is M adjusted by its column means.

\tilde{M}' is \tilde{M} transposed (i.e. switch rows and columns).

c is a normalizing constant (often sum of column variances).

For additive, M is often coded as 0/1/2.

For dominance, M is often coded as 0/1.

For epistasis, GRM is often estimated through Hadamard product of the interaction (e.g. $AA = A \circ A$).

1. Biological, exact

$$\tilde{M}_{A,i} = \begin{matrix} -1 - \mu_{A,i} & \text{gg} \\ 0 - \mu_{A,i} & \text{Gg} \\ 1 - \mu_{A,i} & \text{GG} \end{matrix}$$

$$\tilde{M}_{D,i} = \begin{matrix} 0 - \mu_{D,i} \\ 1 - \mu_{D,i} \end{matrix}$$

$$\tilde{M}_{AA,i} = \begin{matrix} -1 - \mu_{AA,i} \\ 0 - \mu_{AA,i} \\ 1 - \mu_{AA,i} \end{matrix}$$

$$K = \frac{\tilde{M} \cdot \tilde{M}'}{c}$$

2. Biological, approx.

$M_{A,i}$ and $M_{D,i}$ are the same as 1.

$$K = \frac{\tilde{M} \cdot \tilde{M}'}{c}$$

For AA, AD and DD:

$$K_{AA} = K_A \circ K_A - CP(\tilde{M}_A \circ \tilde{M}_A)$$

$$K_{AD} = K_A \circ K_D - CP(\tilde{M}_A \circ \tilde{M}_D)$$

$$K_{DD} = K_D \circ K_D - CP(\tilde{M}_D \circ \tilde{M}_D)$$

3. Statistical, G2A

$$\tilde{M}_{A,i} = \begin{matrix} 0 - 2p_i \\ 1 - 2p_i \\ 2 - 2p_i \end{matrix}$$

$$\tilde{M}_{D,i} = \begin{matrix} -2p_i^2 \\ 2p_i(1 - p_i) \\ -2(1 - p_i)^2 \end{matrix}$$

K_{AA} , K_{AD} and K_{DD} are the same as 2.

4. Statistical, NOIA

$$\tilde{M}_{A,i} = \begin{matrix} -(2 - p_{Gg,i} - 2p_{gg,i}) \\ -(1 - p_{Gg,i} - 2p_{gg,i}) \\ -(0 - p_{Gg,i} - 2p_{gg,i}) \end{matrix}$$

$$\tilde{M}_{D,i} = \begin{matrix} -2p_{Gg,i}p_{gg,i}/x \\ 4p_{GG,i}p_{gg,i}/x \\ -2p_{Gg,i}p_{GG,i}/x \end{matrix}$$

$$x = p_{GG,i} + p_{gg,i} - (p_{GG,i} - p_{gg,i})^2$$

K_{AA} , K_{AD} and K_{DD} are the same as 2.

OK, in simpler terms...

1. Biological, exact

This is the painful way, gets increasingly painful with more markers.

2. Biological, approx.

This is what we normally do.

3. Statistical, G2A

This is known as General 2 Allele (Zeng et al 2005).

Assumes Hardy-Weinberg Equilibrium (HWE) and Linkage Equilibrium (LE).

4. Statistical, NOIA

NOIA: Natural and Orthogonal Interactions.

Assumes only LE.

What is biological vs statistical effect?

$$y = Wb$$

y is the sum of biological effect of 9 genotypes (2 loci).

W is a 9 x 9 matrix of mean and adjusted contrasts.

b is the statistical effect.

A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis.

Alvarez-Castro and Carlborg (2007)

<https://doi.org/10.1534/genetics.106.067348>



Variance components

sim04: background (A, D), main (A, D)



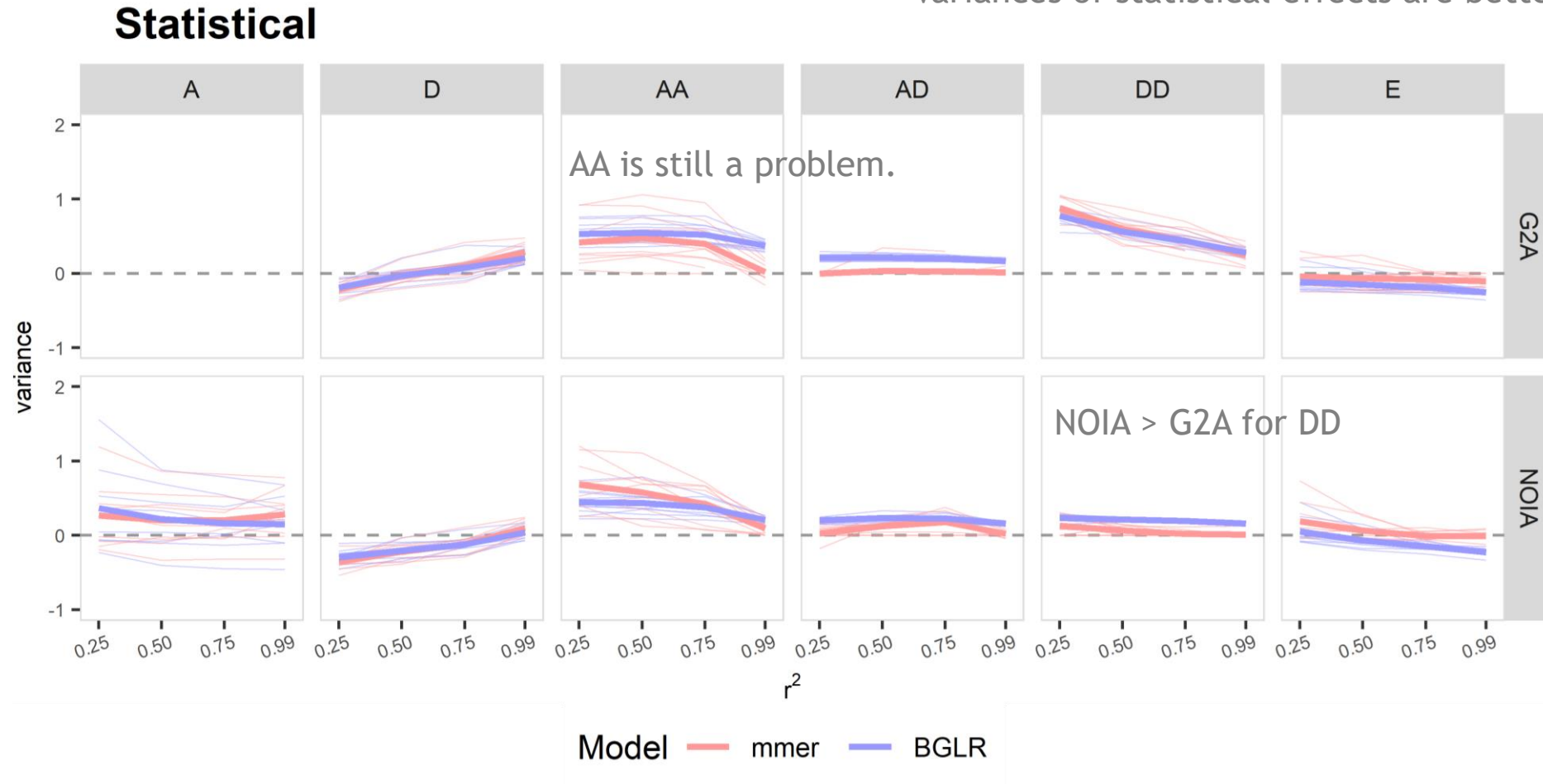
Estimates get worse as marker density drops.

Plot shows the difference between estimated and true values.

Variance components

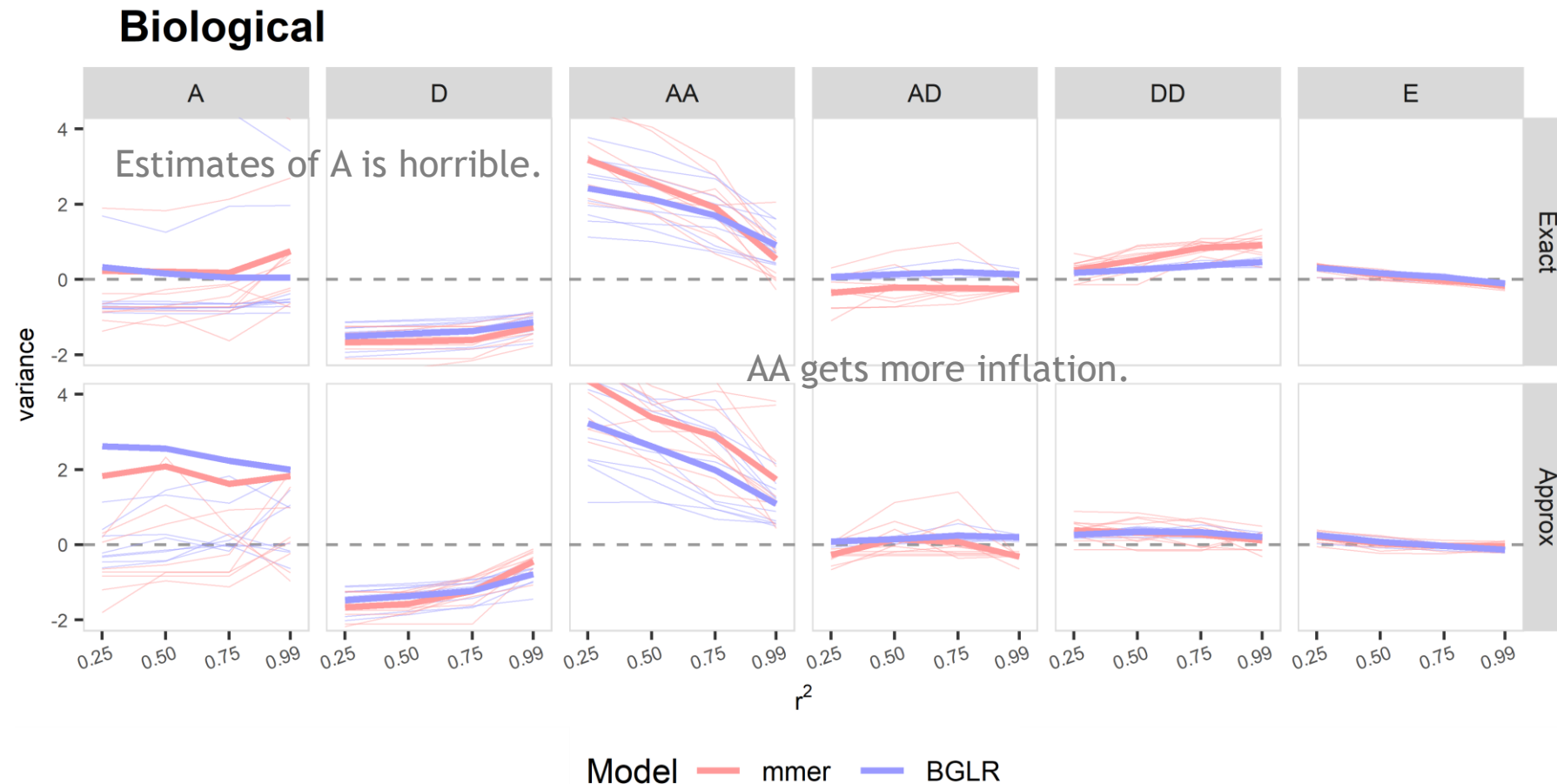
sim04: background (A, D), main (A, D)

Variances of statistical effects are better estimated?



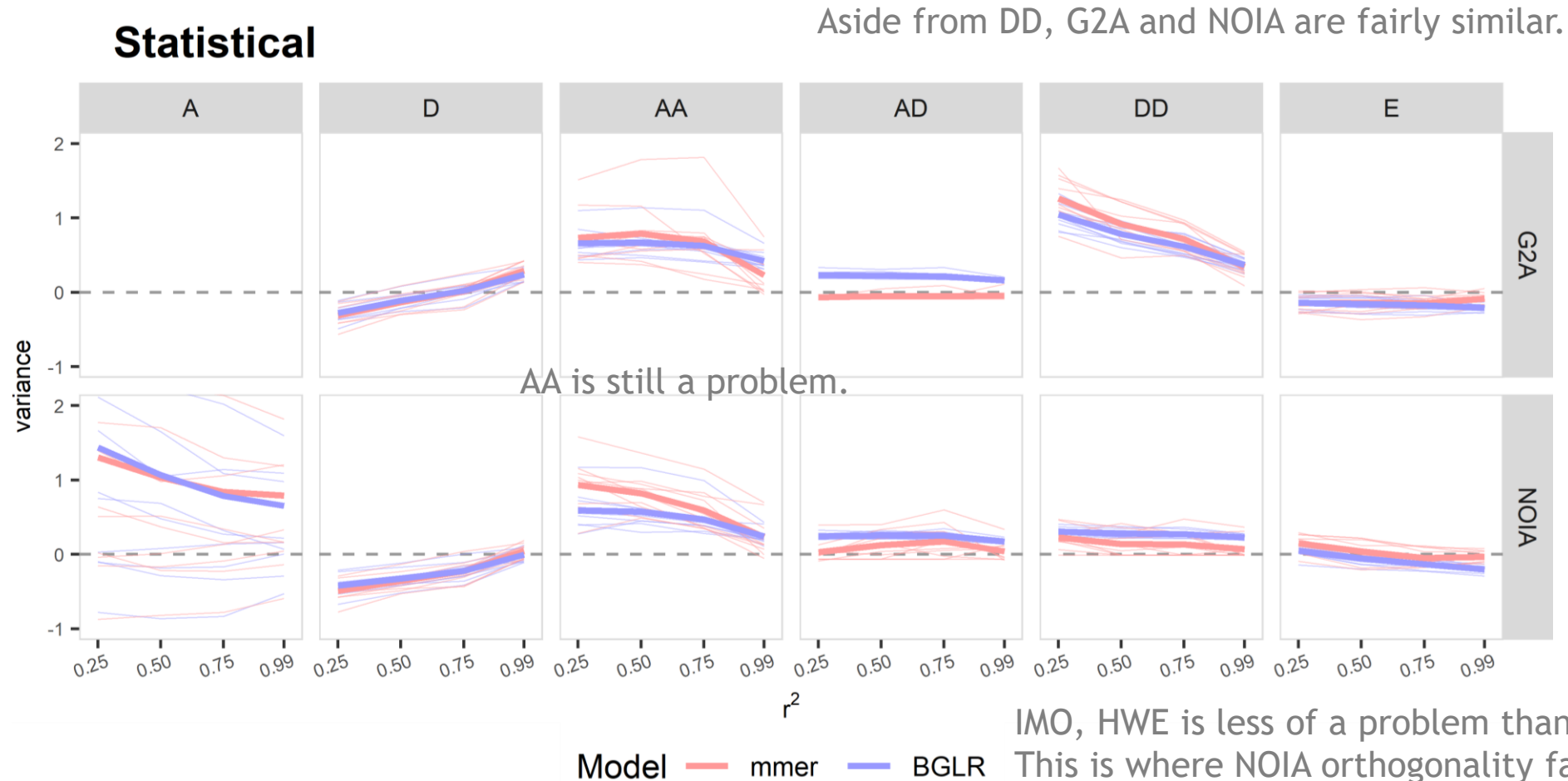
Variance components

sim06: background (A, D, AA, AD, DD), main (A, D)



Variance components

sim06: background (A, D, AA, AD, DD), main (A, D)



Can we use variance components to infer genetic architecture?

Don't know how close we are to the QTLs, and phantom epistasis is pervasive!

The genetic architecture of quantitative traits cannot be inferred from variance component analysis.

Huang and Mackay (2016)

<https://doi.org/10.1371/journal.pgen.1006421>

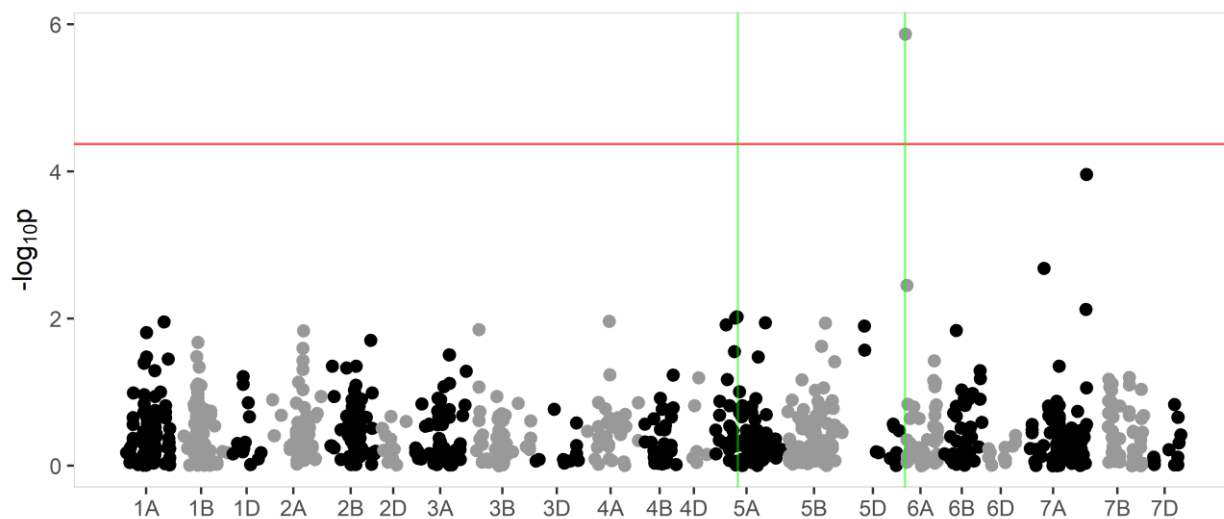


If we are willing to close an eye, estimating statistical effect is still misleading when it comes to inferring the genetic architecture of heterosis.

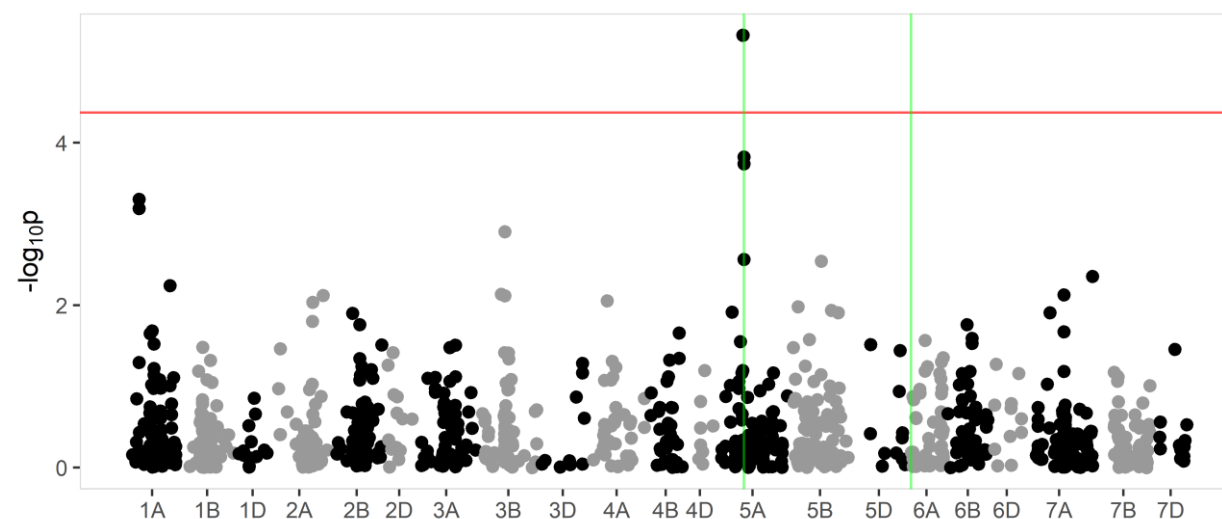
Is GWAS any better?

sim06: background (A, D, AA, AD, DD), main (A, D)

GWAS-A



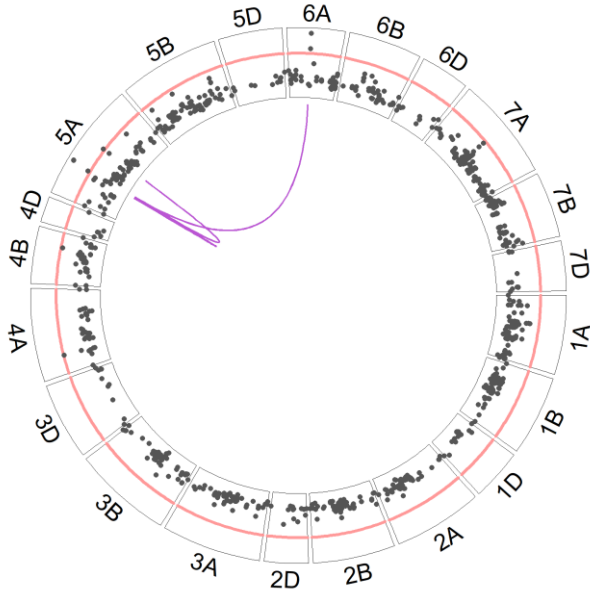
GWAS-D



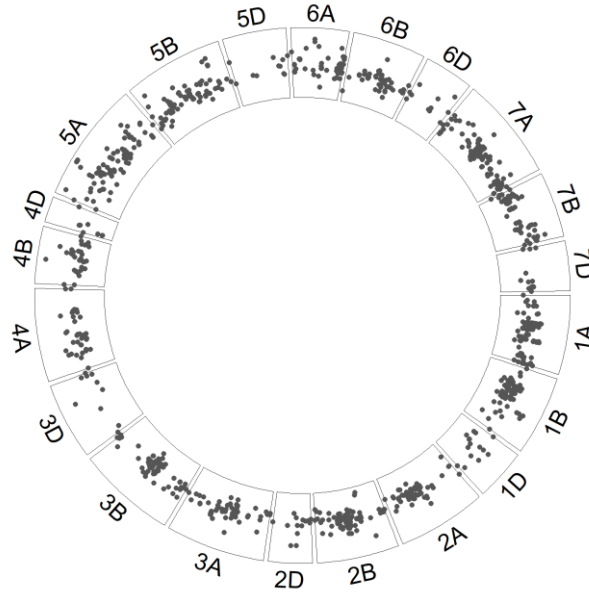
Dataset: LD threshold of $r^2 < 0.5$.

sim06: background (A, D, AA, AD, DD), main (A, D)

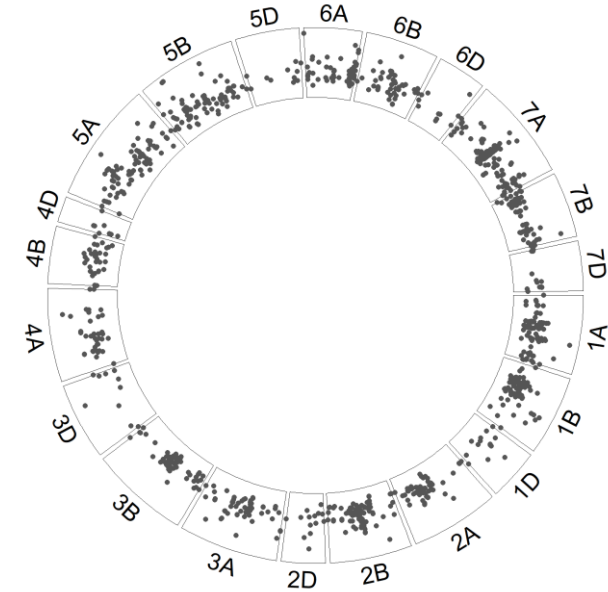
GWAS-AA



GWAS-AD



GWAS-DD

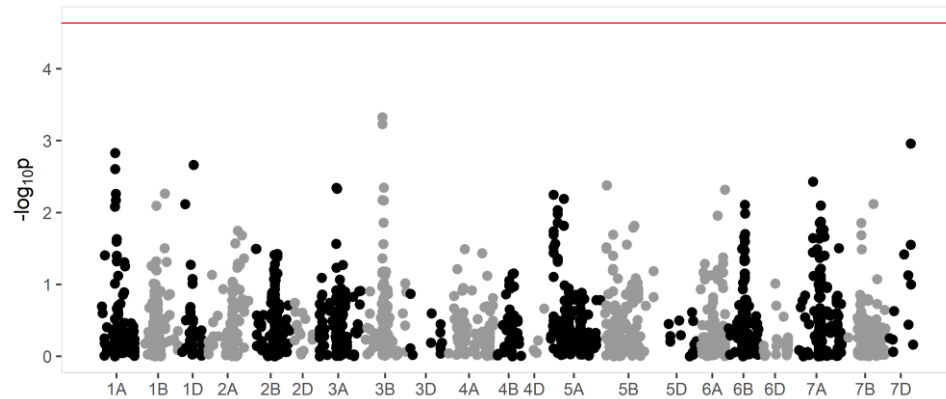


Dataset: LD threshold of $r^2 < 0.5$.

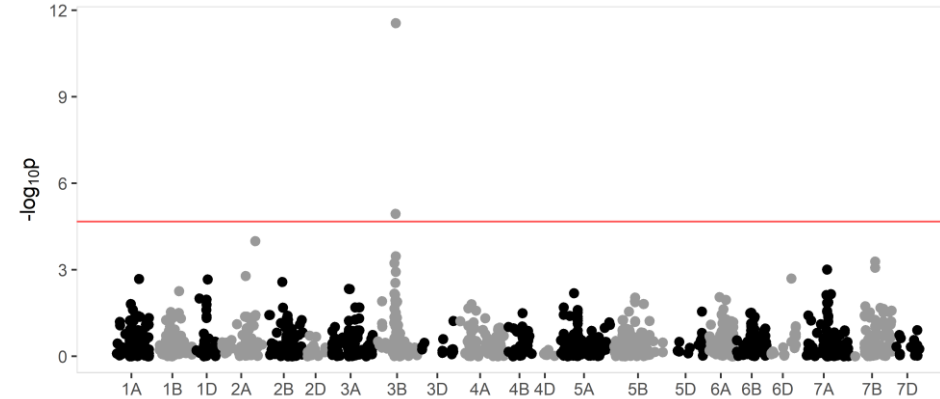
GWAS in simulated traits

GWAS is still running (slowly) - here are some of the previous results (slightly different simulation).

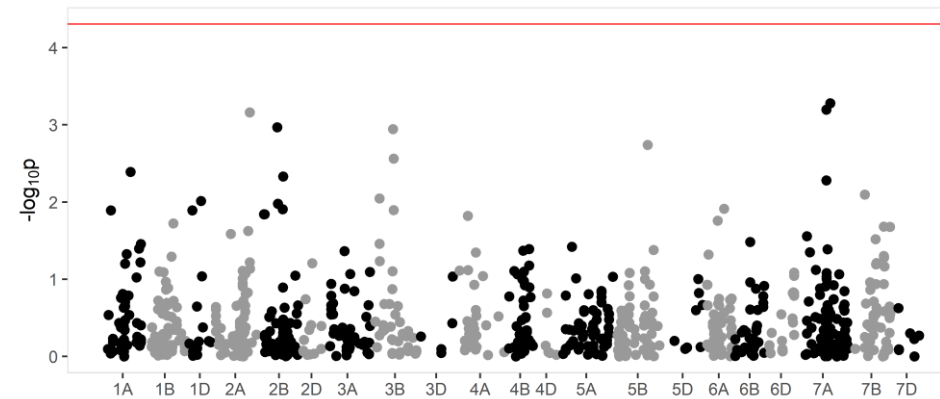
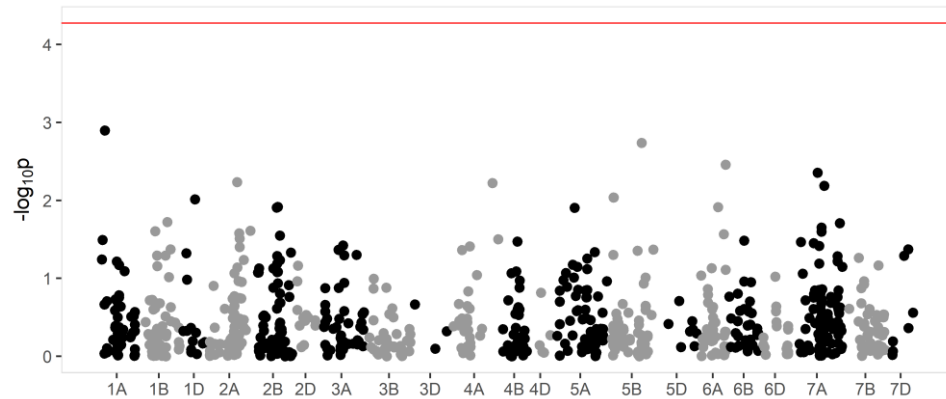
GWAS - Additive



GWAS - Dominance

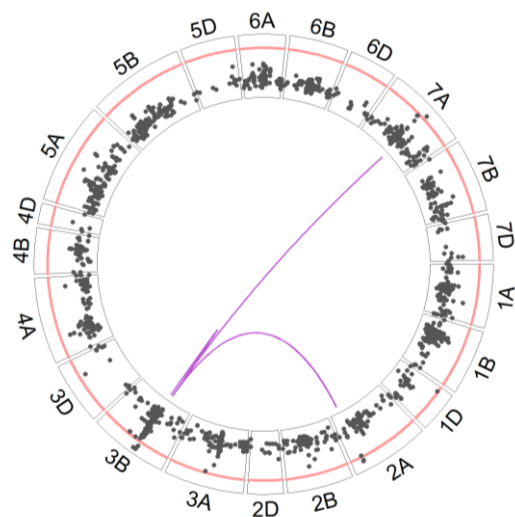


No markers
with $r^2 > 0.99$.

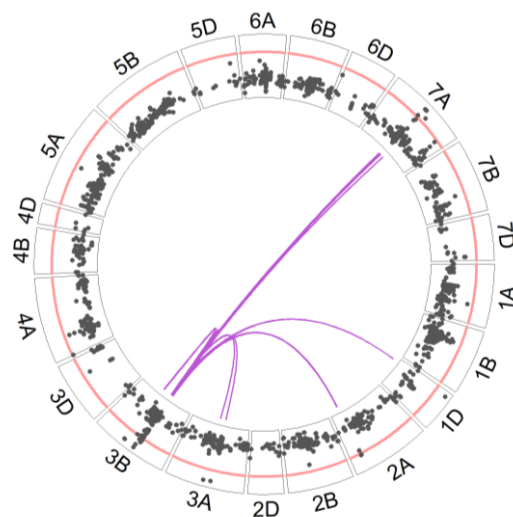


No markers
with $r^2 > 0.25$.

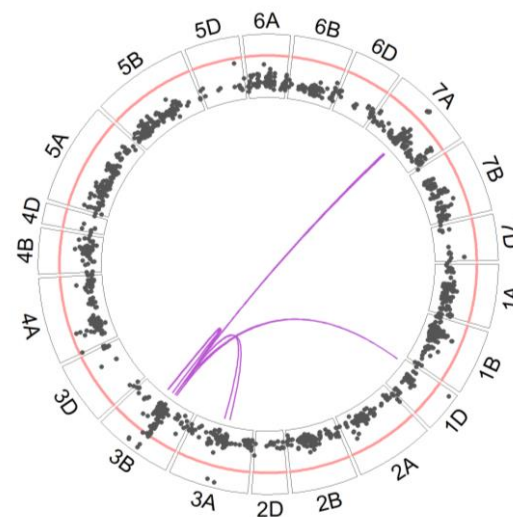
GWAS - A x A



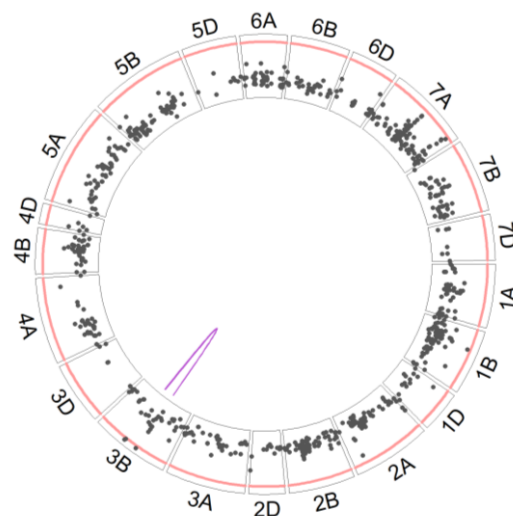
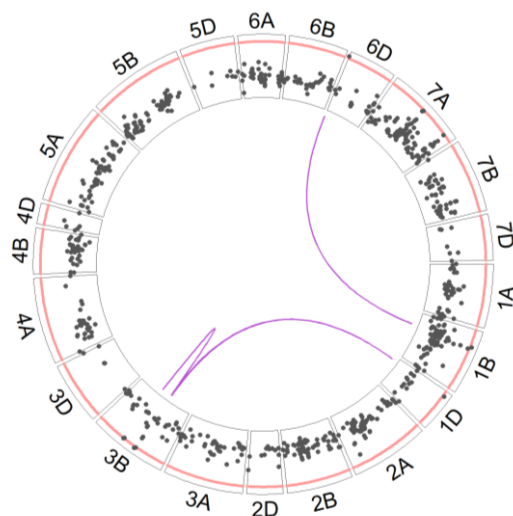
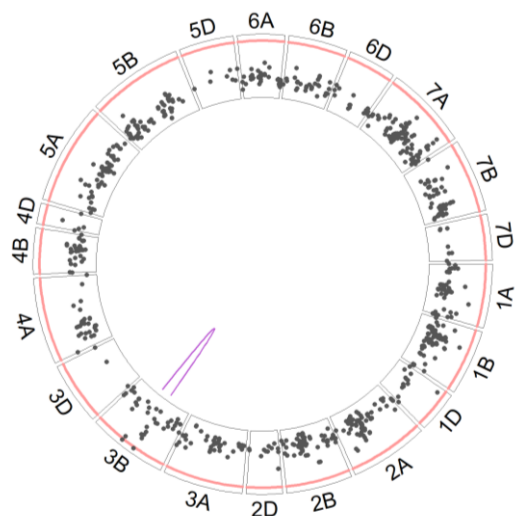
GWAS - A x D



GWAS - D x D



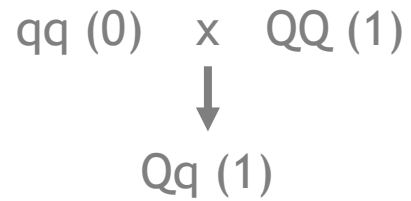
No markers
with $r^2 > 0.99$.



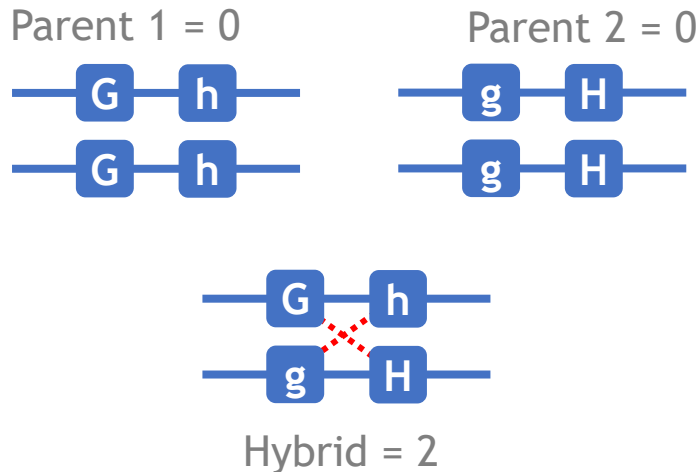
No markers
with $r^2 > 0.25$.

Take home messages

1. Heterosis arises due to directional dominance (and dominance x dominance).

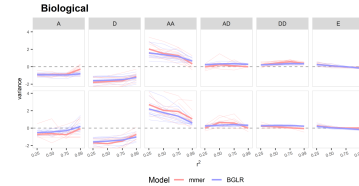
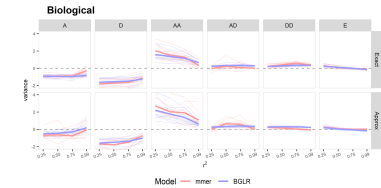


2. Heterosis can also arise due to linked (dispersed) and directional additive x additive interaction.



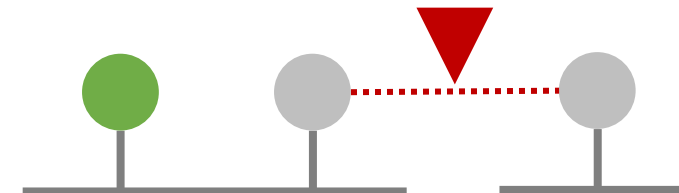
3. Inference of genetic architecture from variance components is hard - further complicated by marker density.

$$V_A = 1.2, V_D = 1.8 \rightarrow$$



✗ $V_D = ?, V_{AA} = ?$

4. Inference of genetic architecture from GWAS is not any better - hard to control for false positive in interaction GWAS.



Principal's Research Group

The Principal's Research Group brings together plant breeding expertise from across SRUC to boost the productivity, sustainability and resilience of food systems across the world.

Wayne Powell

Ian Mackay

Rajiv Sharma

Ian Dawson

David Marshall

Nicola Rossi

Please feel free to reach out to us if you have any question or interest in collaboration!



@hataraku_cj

@SRUCPrincipal

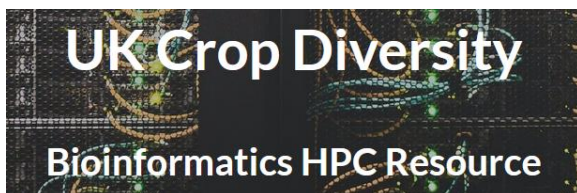
@IanJMackay

cyang@sruc.ac.uk

Many thanks to various authors for making their data publicly available.

And very grateful to Jon and CGDG for the invite.

External computational resource



Slides are available at: <https://github.com/cjyang-work/presentation/>



Microbe-dependent trait expression in wheat.

<https://findaphd.com/search/ProjectDetails.aspx?PJID=148313>

**Due on the 5th of December.
There is still time to come
and join our group!**



Developing a commercial breeding program to support the cultivation of the red seaweed dulse (*Palmaria palmata*)

<https://findaphd.com/search/ProjectDetails.aspx?PJID=148310>



Diversity, genomic characterisation and simulation of breeding potential for a new food crop for Scotland/UK.

<https://findaphd.com/search/ProjectDetails.aspx?PJID=148283>



Biochemical, metabolomic & agronomic profiling of the tuberous legume food crop, *Apios americana*, for Scotland/UK.

<https://findaphd.com/search/ProjectDetails.aspx?PJID=148280>



Rapid domestication of purslane (*Portulaca* sp.) in a vertical farm environment.

<https://findaphd.com/search/ProjectDetails.aspx?PJID=148270>