Sequence-based genomic prediction for a complex trait in *Drosophila melanogaster* reveals sex-differentiated epistasis

Henner Simianer[1], Ulrike Ober[1], Wen Huang[2] & Trudy Mackay[2]

[1] Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August-University Göttingen, Germany

[2] Department of Genetics, North Carolina State University, Raleigh, USA





Drosophila melanogaster Genetics Reference Panel (DGRP)

- 176 inbred lines
- for each line ~ 100 males and 100 females phenotyped
- all lines fully sequenced with ~2.5 mio SNPs
- Genomic prediction with GBLUP
- accuracy of prediction evaluated by 5-fold cross-validation (CV)

	CV accuracy			H^2
	all	only males	only females	
Starvation resistance	0.24	0.20	0.25	0.59
Startle response	0.23	0.23	0.22	0.57
Chill coma recovery	-.04	-.14	0.05	0.37

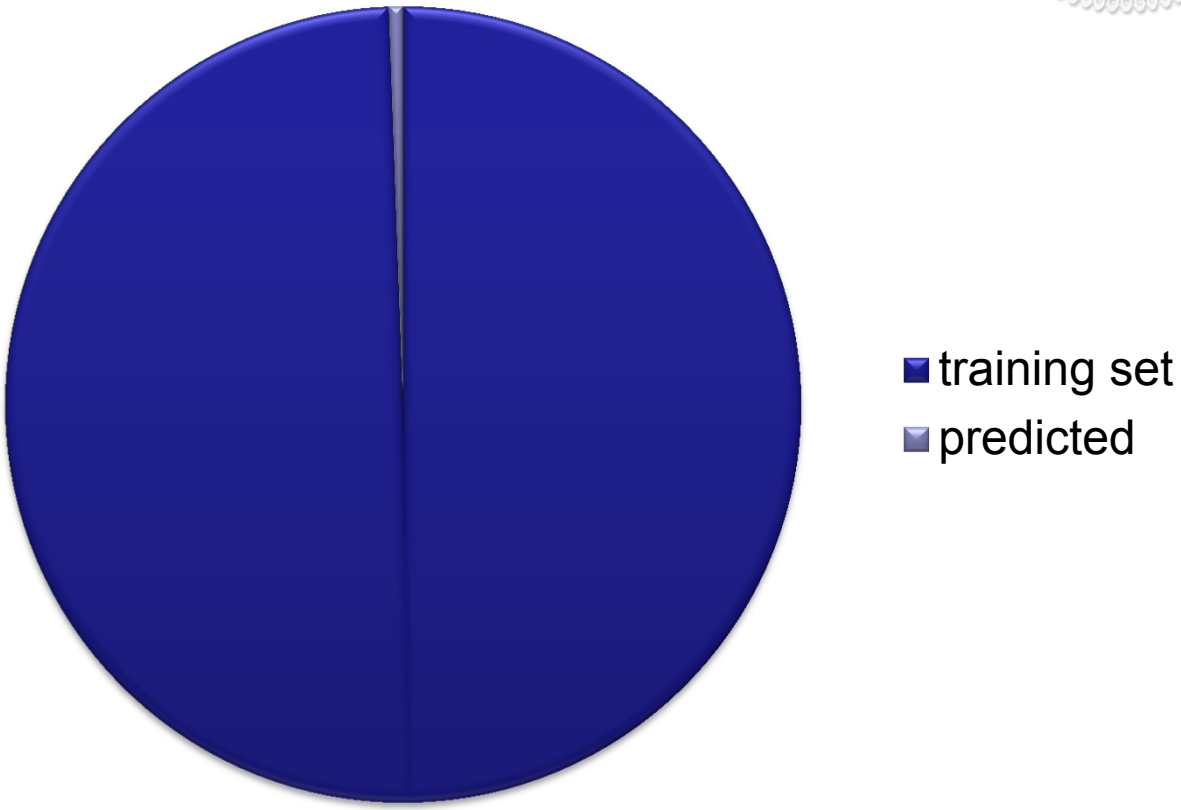
Ober *et al.* (2012) PLoS Genetics

Why does genomic prediction fail for the heritable trait chill coma recovery while it works for other traits?

Obvious candidate reasons (non-normal distribution, outliers etc.) could be ruled out

Leave-one-out cross-validation

- ⇨ 175 lines in training set
- ⇨ 1 line predicted
- ⇨ 176 replicates



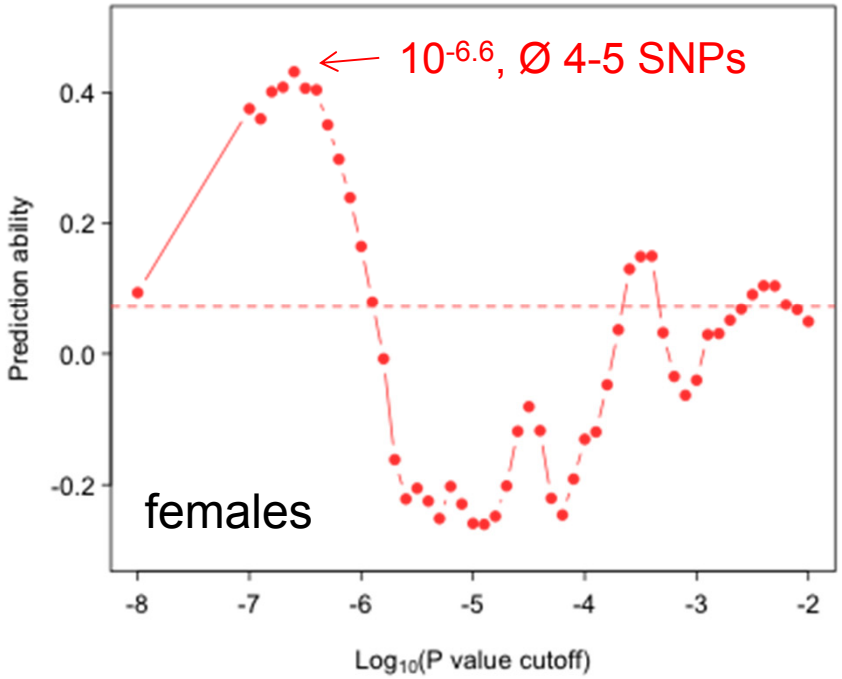
Obtained accuracy with all SNPs:

males: NA (in most cases $\hat{\sigma}_g^2 = 0$)

females: 0.059

Poor man‘s Bayes B

- ⇨ 1,868,905 common variants (MAF >= 0.05)
- ⇨ 175 lines in training set
  - ⇨ GWAS in the training set
  - ⇨ select all SNPs with $p < 10^{-x}$
- ⇨ predict remaining line just with this subset of SNPs
- ⇨ repeat 176 times so that each line is predicted once



How to include additive x additive epistasis

Additive genomic relationship matrix (VanRaden, 2008)

Matrix M: # individuals x # genotypes, coded as -1,(0),1

Matrix P: # individuals x # genotypes, column i is $2 \cdot (p_i - 0.5)$

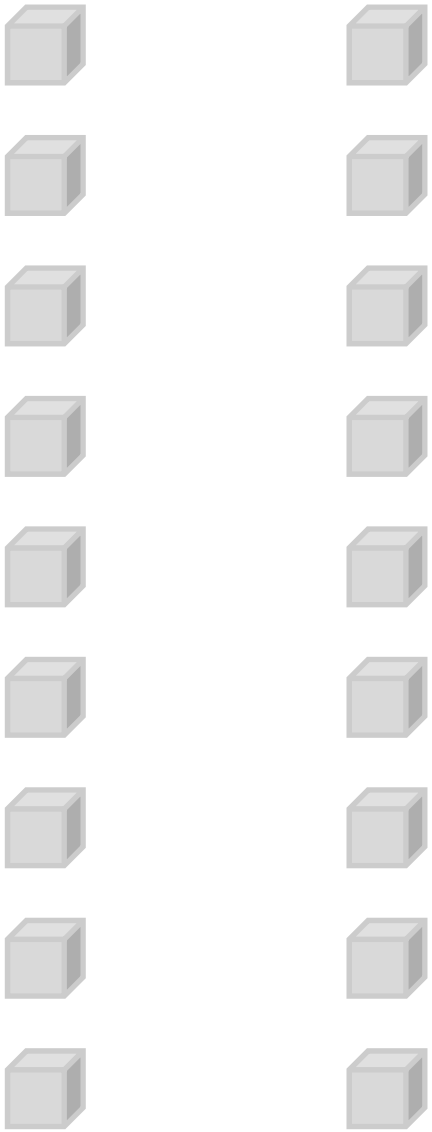
$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{2 \cdot \sum_{i=1}^{n_{SNPs}} (p_i \cdot (1 - p_i))}$$

$$\mathbf{G}_{AxA} = \mathbf{G} \circ \mathbf{G}$$

Without SNP-selection

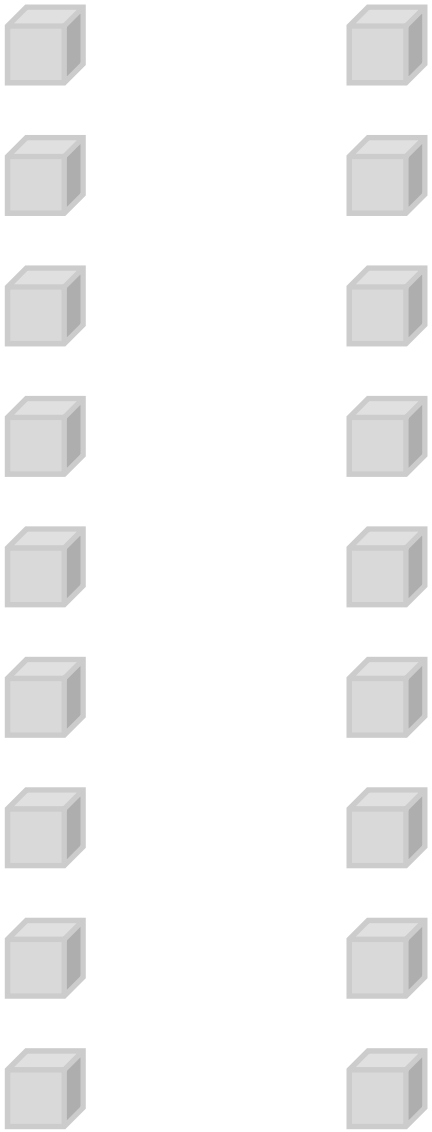
Prediction with the epistatic covariance matrix $\mathbf{G}_{AxA}$ based on all SNPs

⇨ **Prediction ability: ~0**



With SNP-selection

1. Identify significant additive x additive interactions in an epistatic GWAS

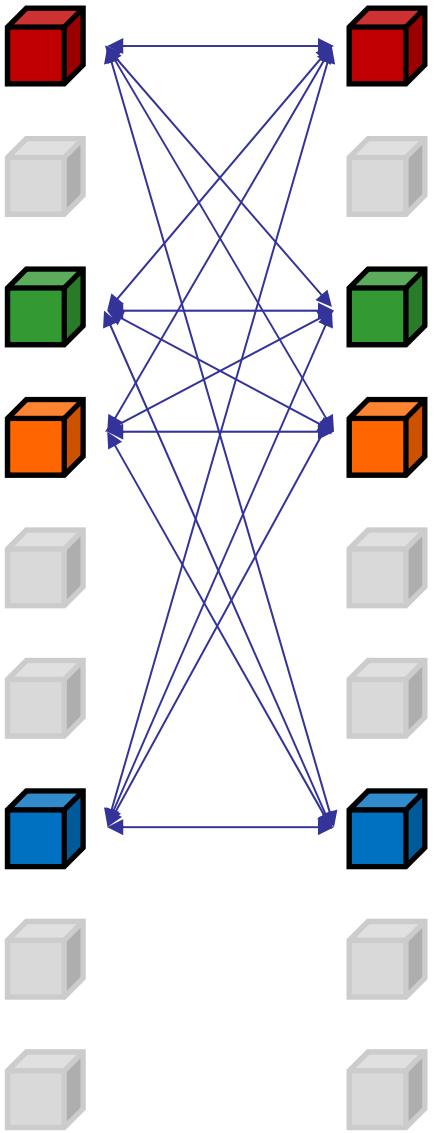


With SNP-selection

1. Identify significant additive x additive interactions in an epistatic GWAS
2. Build the $\mathbf{G}^*$ matrix for just the SNPs included in the pairs
3. Construct the epistatic matrix

$$\mathbf{G}^*_{AxA} = \mathbf{G}^* \circ \mathbf{G}^*$$

⇨ **Prediction ability with this model: ~0**



Statistical Science
2009, Vol. 24, No. 4, 451–471
DOI: 10.1214/09-STS307

Population Structure and Cryptic Relatedness in Genetic Association Studies

William Astle and David J. Balding[1]

$$\mathbf{G}_{AB} = \frac{1}{n_{SNPs}} \sum_{i=1}^{n_{SNPs}} \frac{(\mathbf{m_i} - \mathbf{p_i})(\mathbf{m_i} - \mathbf{p_i})'}{2 \cdot p_i \cdot (1 - p_i)}$$

VanRaden (2008): $$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{2 \cdot \sum_{i=1}^{n_{SNPs}} (p_i \cdot (1 - p_i))}$$

Extention of the Astle & Balding approach for additive x additive epistasis

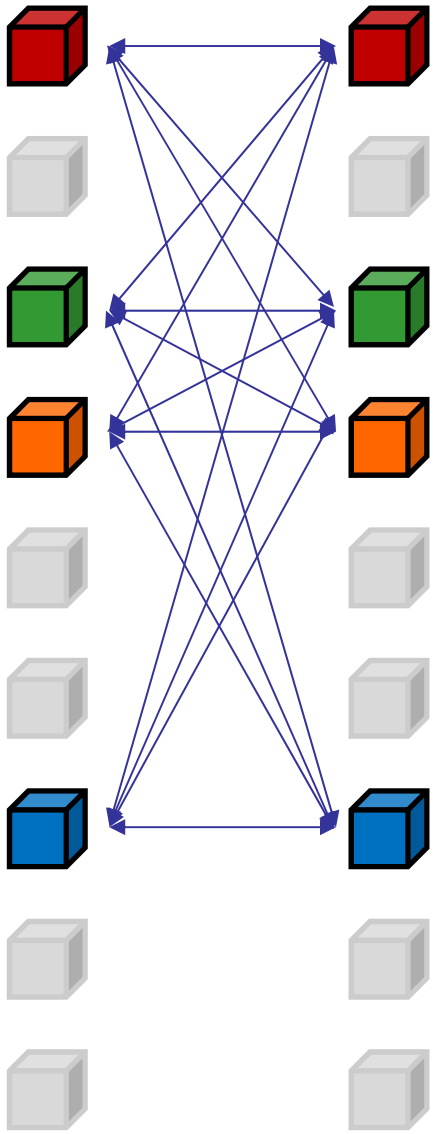
Epistatic GWAS ⇨ $\mathrm{k} = 1, \ldots, n_{EP}$ significant SNP pairs $\{k_1, k_2\}$

Construct a matrix for each SNP $\mathbf{G}_{ki} = \frac{(\mathbf{m}_{ki} - \mathbf{p}_{ki})(\mathbf{m}_{ki} - \mathbf{p}_{ki})'}{2 \cdot p_{ki} \cdot (1 - p_{ki})}$

Then build $\mathbf{G}_{AB_{AxA}} = \frac{1}{n_{EP}} \sum_{k=1}^{n_{EP}} \mathbf{G}_{k1} \circ \mathbf{G}_{k2}$

With SNP-selection

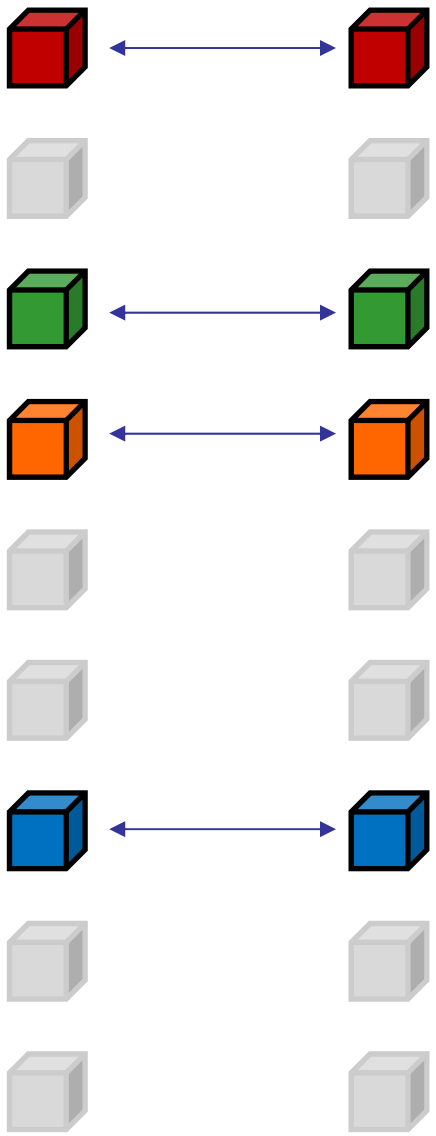
1. Identify significant additive x additive interactions in an epistatic GWAS
2. Build the $\mathbf{G}_{AB_{AxA}}$ matrix with all significant pairs



With SNP-selection

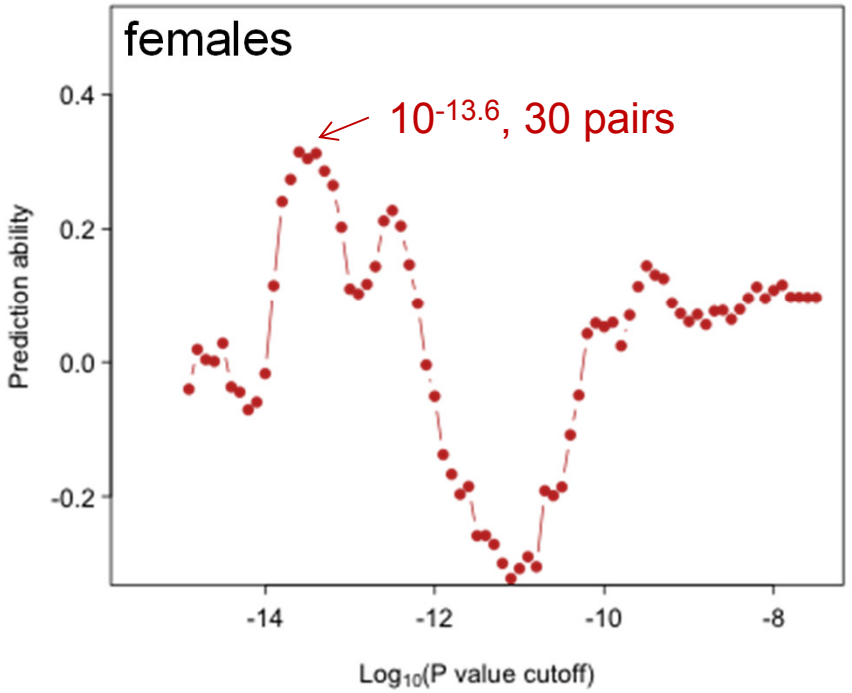
1. Identify significant additive x additive interactions in an epistatic GWAS
2. Build the $\mathbf{G}_{AB_{AxA}}$ matrix with all significant pairs

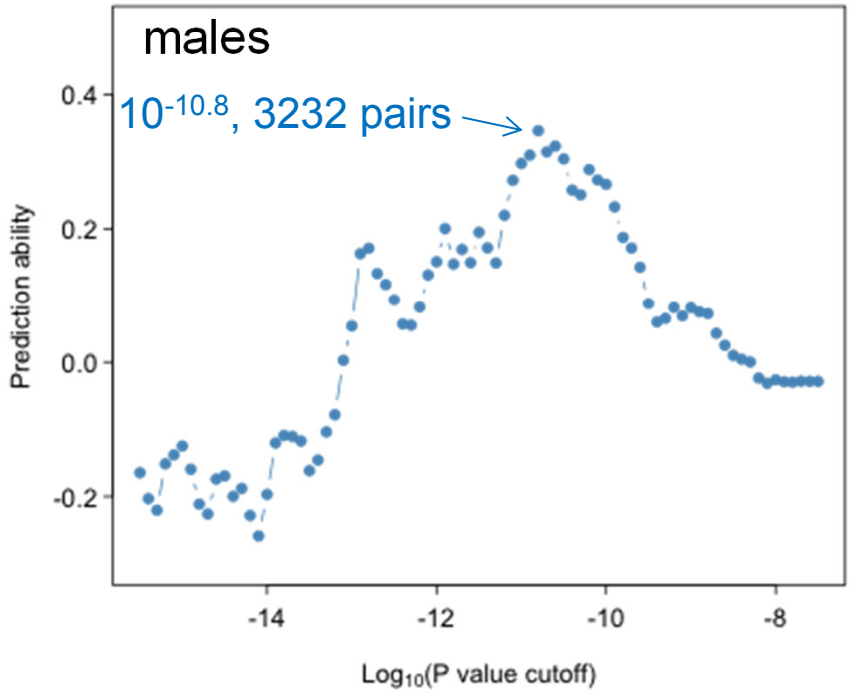
⇨ **Prediction ability with this model ...**



Leave-one-out cross-validation – epistatic SNP selection

- ⇨ 672,636 LD-pruned frequent variants (MAF >= 0.15)
- ⇨ 175 lines in training set
  - ⇨ do an additive x additive GWAS in the training set (2.2×10^{11} pairs)
  - ⇨ construct the $\mathbf{G}_{AB_{AxA}}$ matrix only with those SNP pairs for which $p < 10^{-x}$
  - ⇨ predict the remaining line
- ⇨ repeat this 176 times



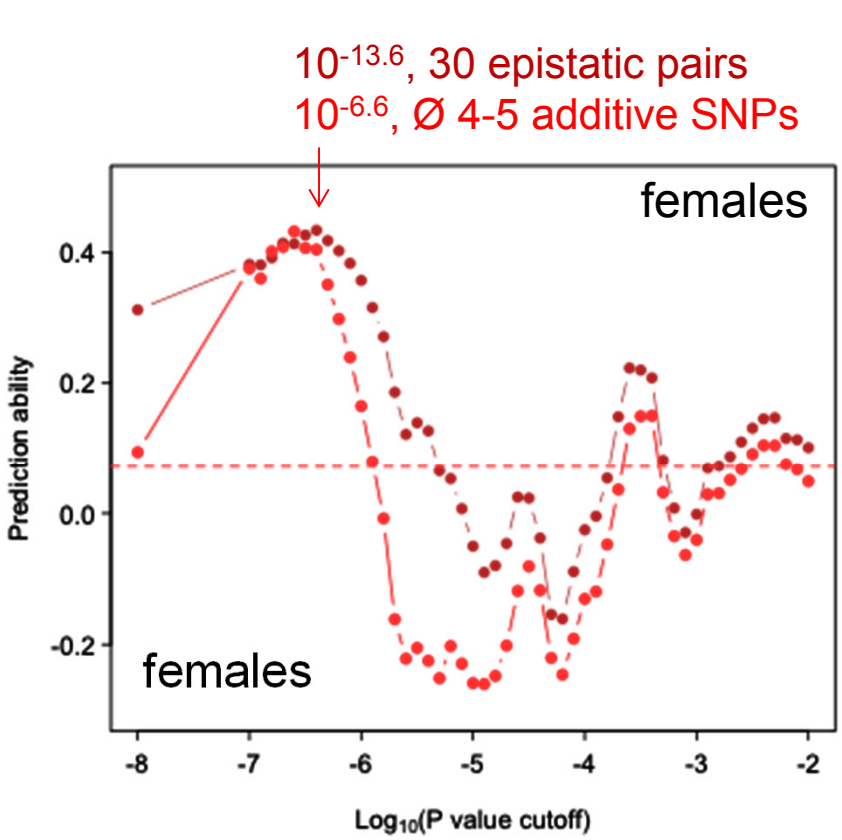


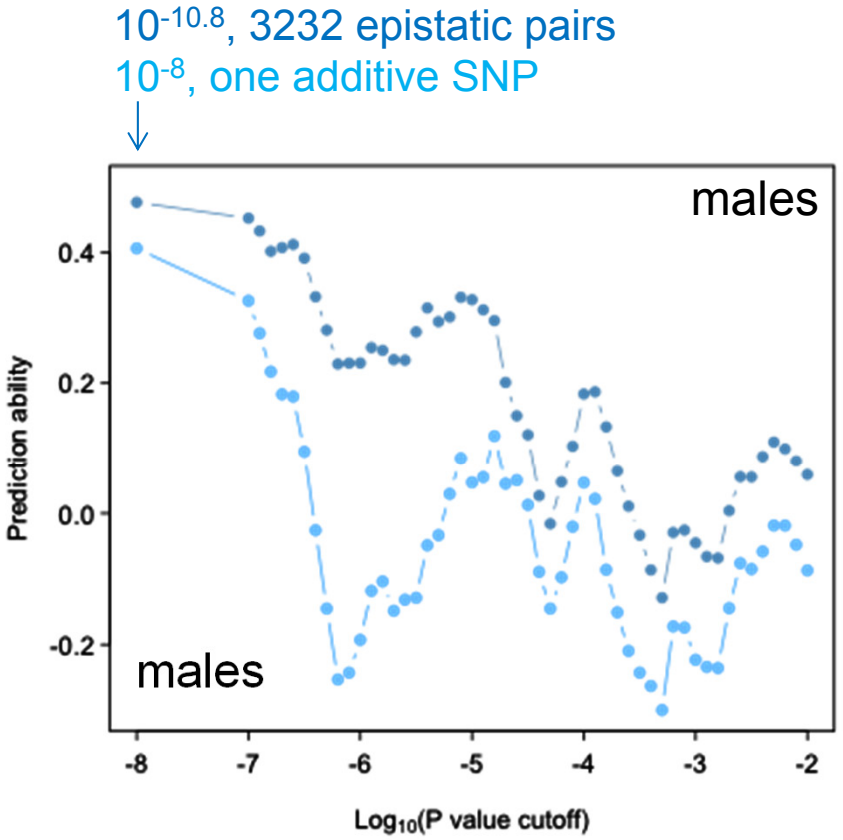
Combined additive + epistatic scan

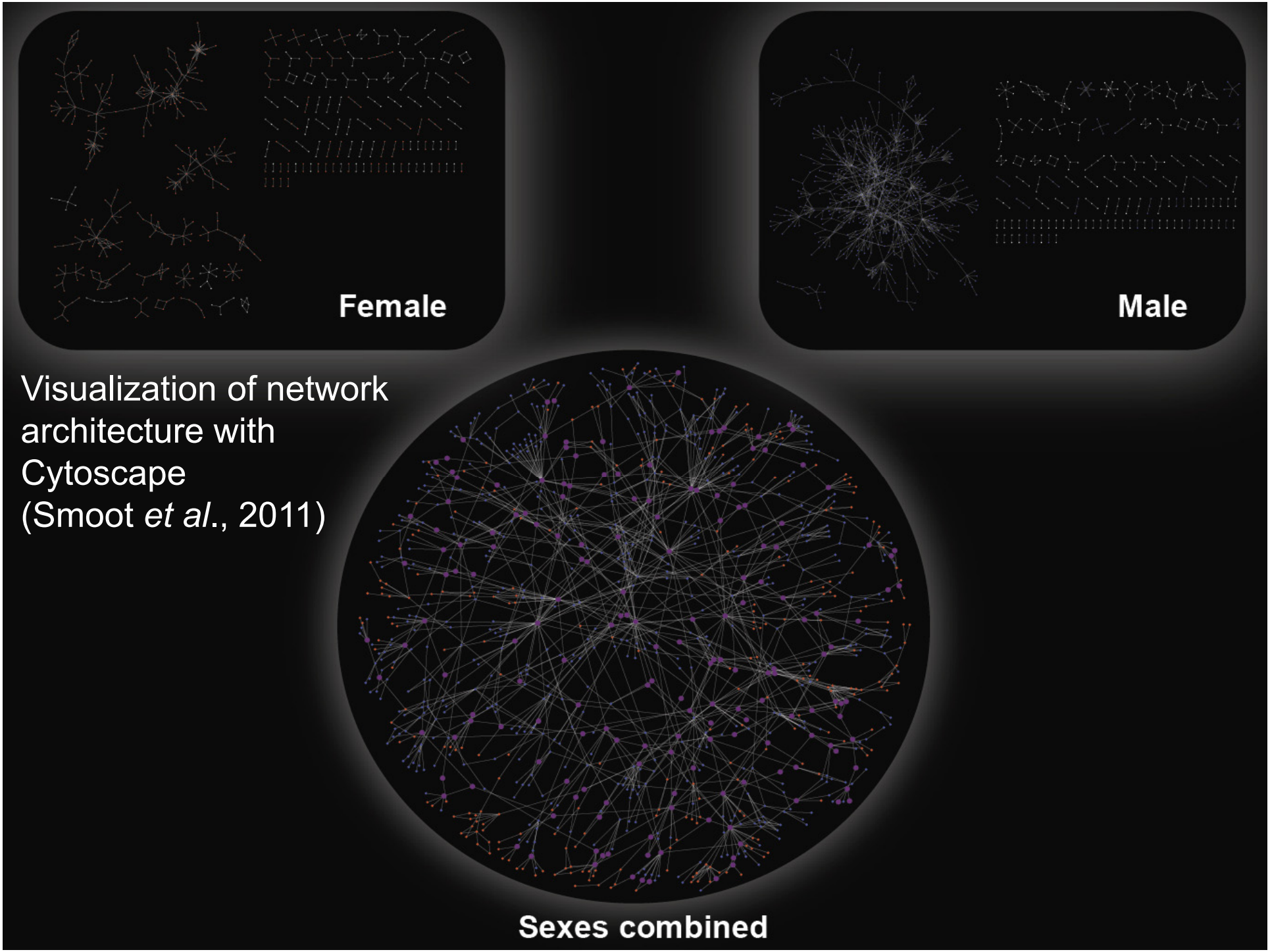
⇨ chose the epistatic set with the highest predictive ability

⇨ add an additive scan across the whole scale

⇨ predict with a combined model (additive + epistatic)







Visualization of network architecture with Cytoscape (Smoot *et al*., 2011)

Summary and conclusions

- Chill coma resistance in *Drosophila melanogaster* is a trait for which genomic prediction with GBLUP fails, although genetic variance exists
- GWAS-based pre-selection of the most significant SNPs improves massively the prediction ability in an additive model
- When properly modeled, epistatic additive x additive interactions also provide a comparable prediction ability
- Combining the top additive and additive x additive effects in the same model yields a prediction ability ~0.4, compared to zero with GBLUP
- The trait chill coma resistance was found to have a rather different genetic architecture in males and females
- Predicting performance of one sex with a model optimized for the other sex essentially failed

What could this result mean for animal breeding?

- Traits expressed in males and females (such as growth-related traits) may have very different genetic architecture (despite having a high genetic correlation, r_{MF} for chill coma resistance was 0.87)
- Genomic prediction relies on SNPs that capture the underlying genetic architecture of a trait (especially so for methods with feature selection such as Bayes B)
- A model trained with male performance data may thus fail to accurately predict female performances (and *vice versa*)
- Empirical validation of this hypothesis needed

Thank you


Ulrike

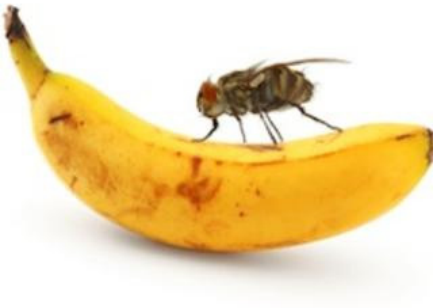

Wen


Trudy

This research was funded by the German Federal Ministry of Education and Research within the AgroClustEr “Synbreed – Synergistic plant and animal breeding” (Funding ID: 0315528C) in association with the DFG research training group ”Scaling problems in statistics” (RTG 1644).

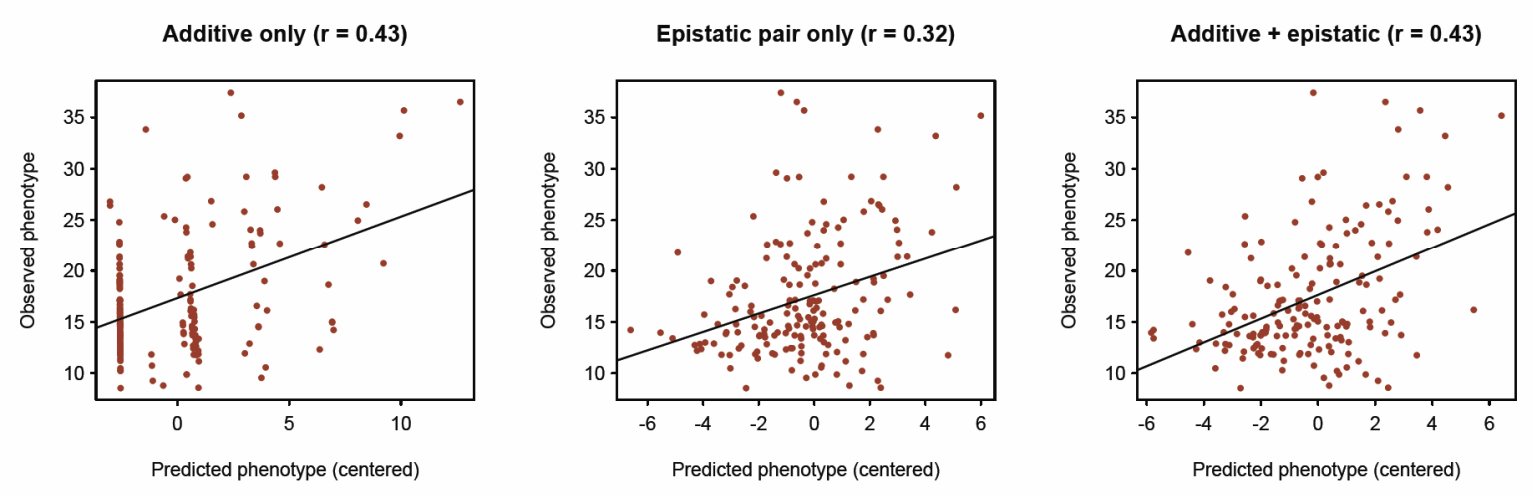
DFG Deutsche Forschungsgemeinschaft

Synbreed

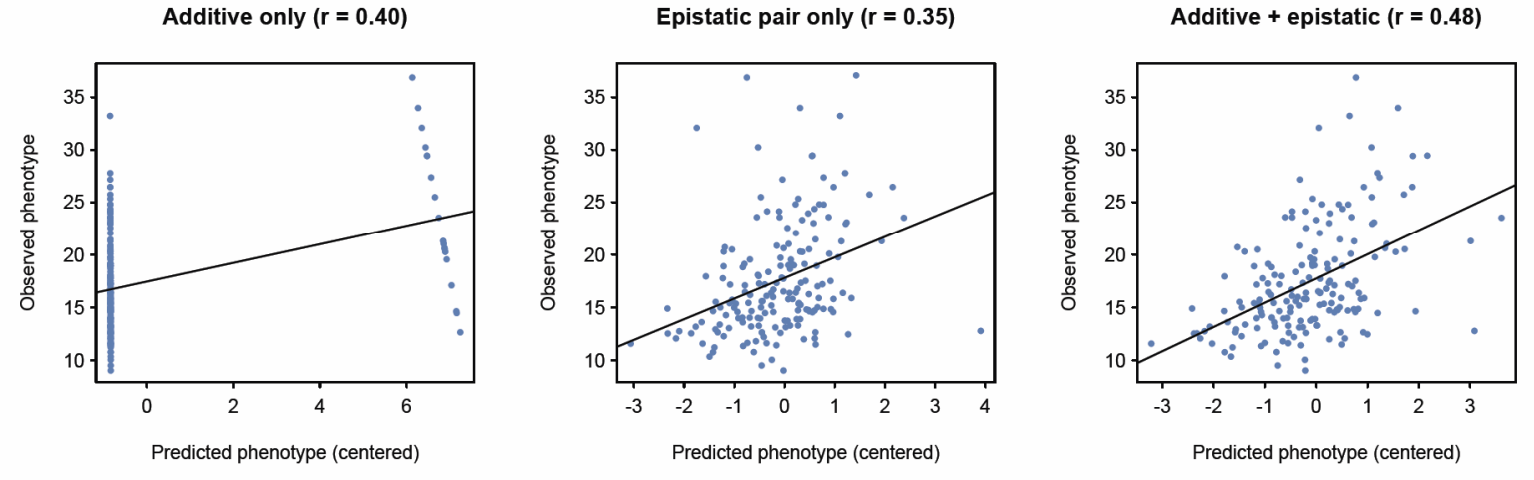


Predicted vs. observed phenotypes with the optimal model in the leave-one-out crossvalidation

females

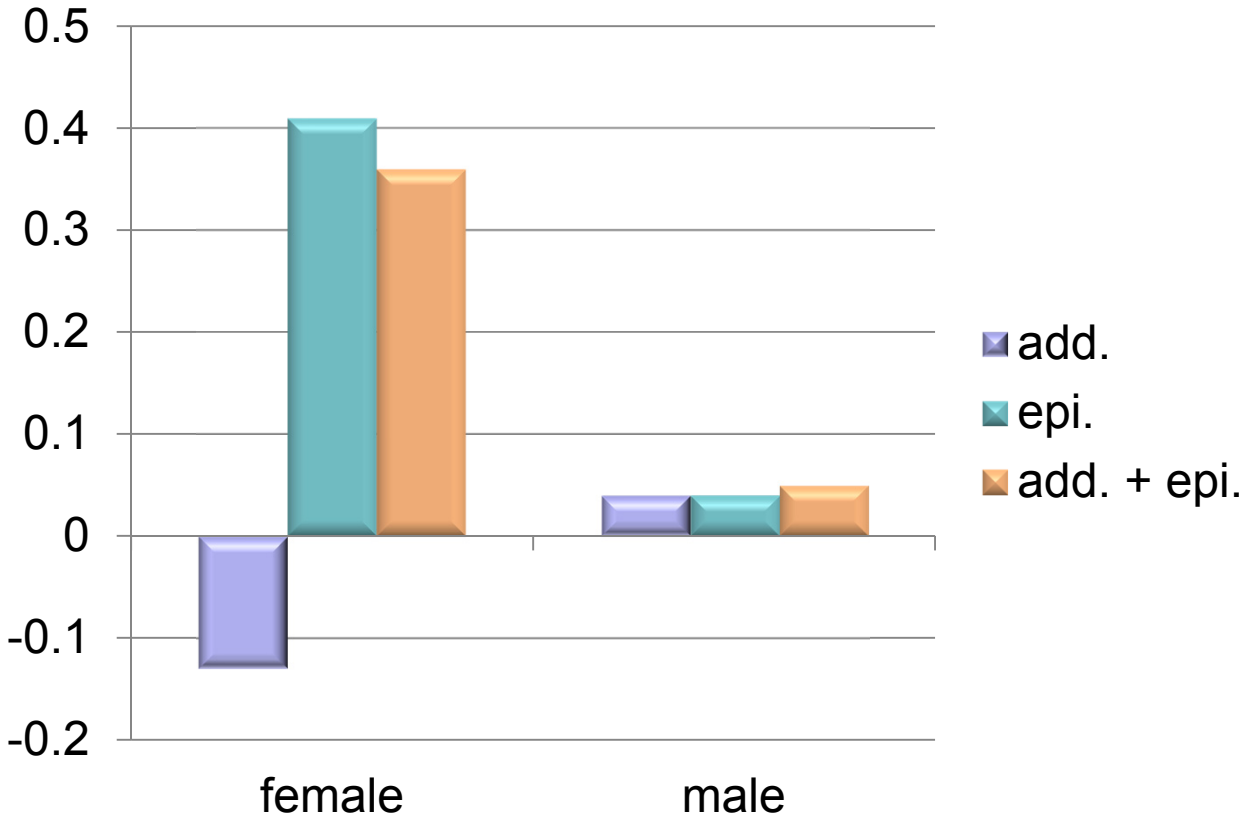


males



The proof of the pudding ...

External validation by predicting an additional set of 27 lines sequenced and phenotyped (~50 replicates per line and sex) in 7/2013



ANOVA with individual measurements
(176 lines × 200 individuals ≈ 35‘000 measurements)

$$\text{phenotype} = \mu + \text{sex} + \text{line} + \text{line} * \text{sex} + \text{replicate}(\text{sex} * \text{line}) + \text{residual} \qquad \text{(Model 1)}$$

$$line \sim N(0, \sigma_l^2 I)$$

$$\text{phenotype} = \mu + \text{sex} + \text{line} + \text{line} * \text{sex} + \text{replicate}(\text{sex} * \text{line}) + g + \text{residual} \qquad \text{(Model 2)}$$

$$g \sim N(0, \sigma_g^2 G)$$

$$\text{phenotype} = \mu + \text{sex} + \text{line} + \text{line} * \text{sex} + \text{replicate}(\text{sex} * \text{line}) + g + (g \times g) + \text{residual} \qquad \text{(Model 3)}$$

$$g \times g \sim N(0, \sigma_{g \times g}^2 G \circ G)$$

Variance components obtained with ASREML

Starvation resistance

	σ_l^2	σ_g^2	$\sigma_{g\times g}^2$	σ_e^2
Model 1	88.0	-	-	88.0
Model 2	0	43.1	-	
Model 3	0	43.1	0	

Startle response

	σ_l^2	σ_g^2	$\sigma_{g\times g}^2$	σ_e^2
Model 1	33.5	-	-	25.7
Model 2	0	16.5	-	
Model 3	0	13.0	1.7	

Chill coma recovery

	σ_l^2	σ_g^2	$\sigma_{g\times g}^2$	σ_e^2
Model 1	23.4	-	-	50.2
Model 2	19.8	1.8	-	
Model 3	0	0	5.9	

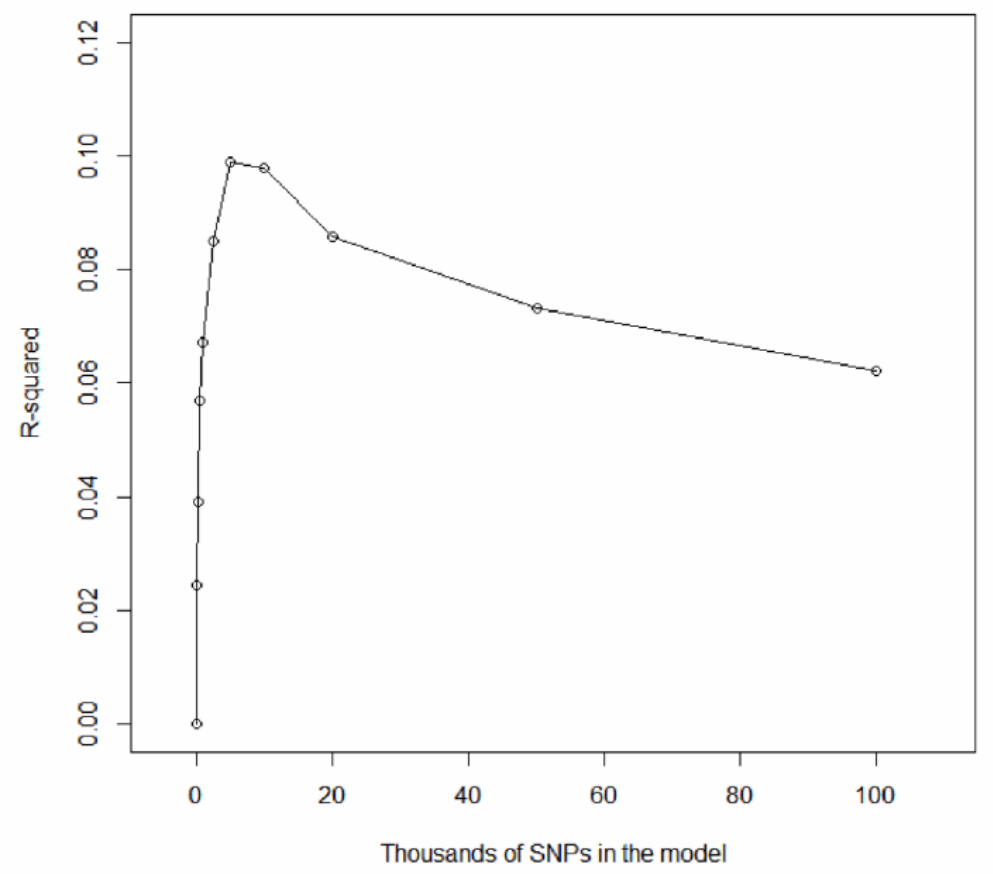
OPEN ACCESS Freely available online

PLOS | GENETICS

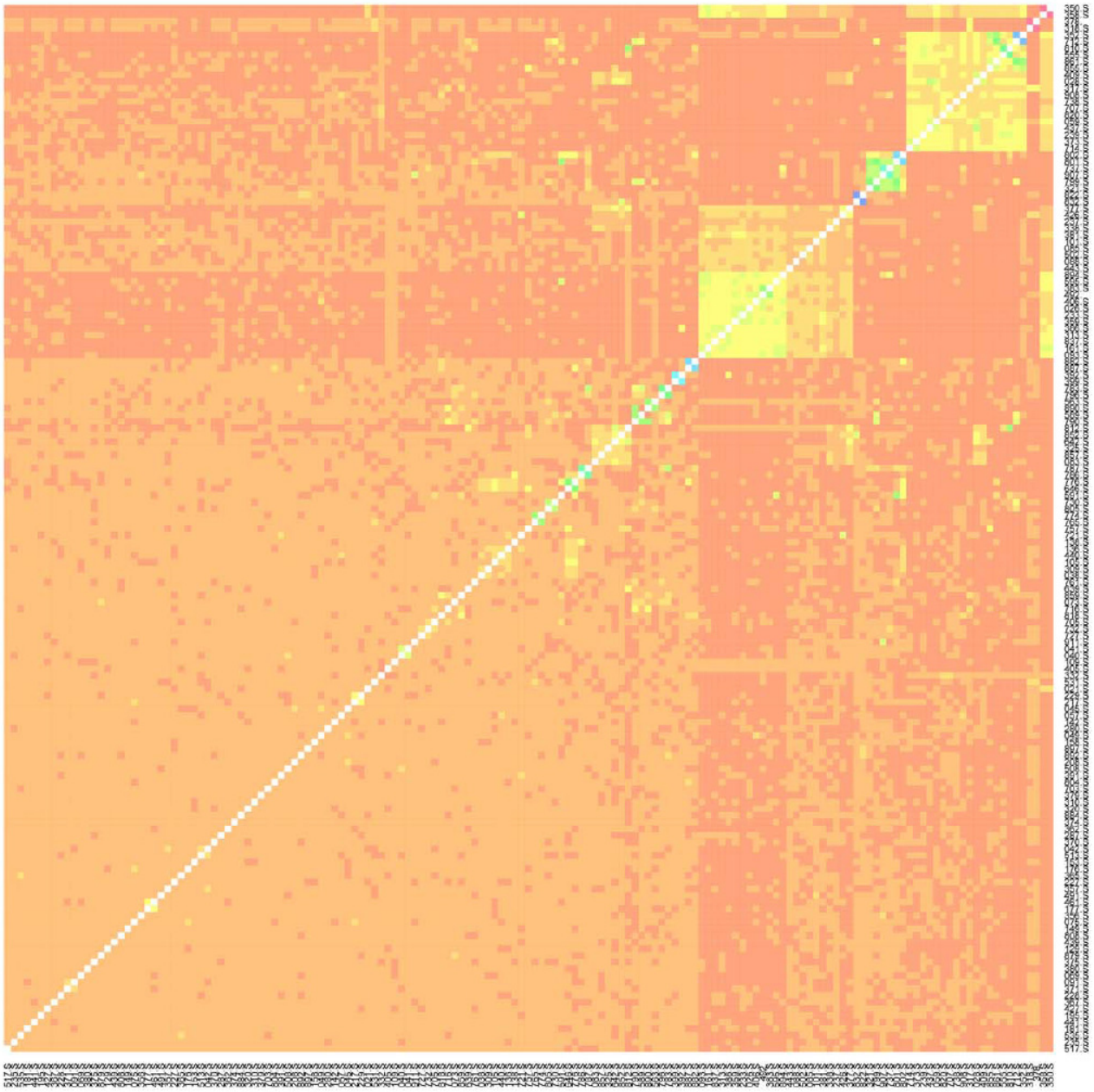
Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor

Gustavo de los Campos[1]*, Ana I. Vazquez[1], Rohan Fernando[2], Yann C. Klimentidis[3], Daniel Sorensen[4]

Genomic prediction in (largely) unrelated samples gains from constructing the G matrix only from the most significant SNPs in a GWAS



Are the DGRP lines largely unrelated?



Heatmap of G