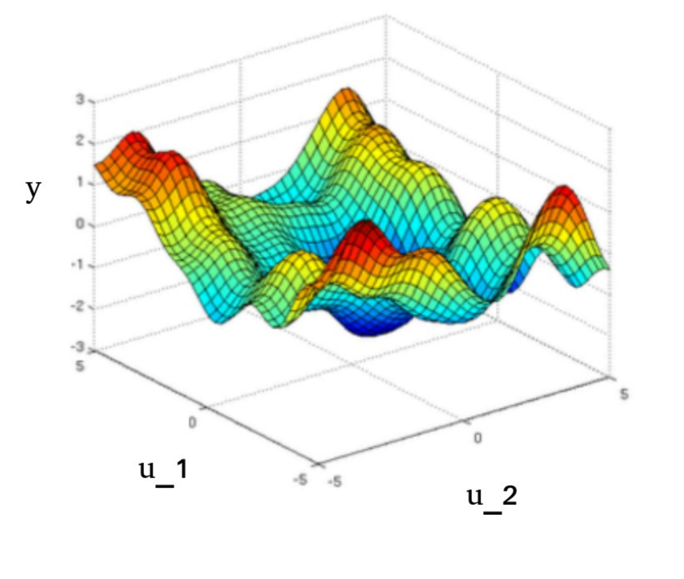
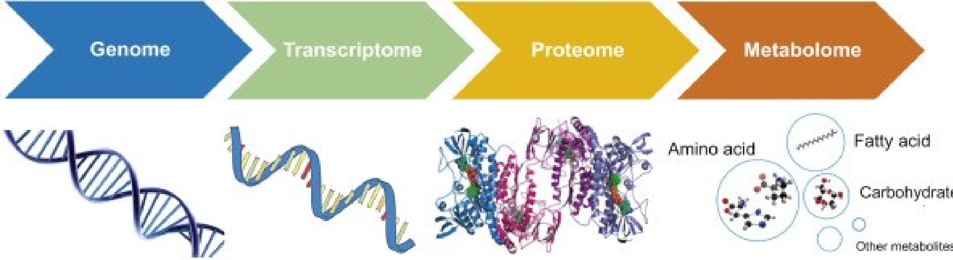
Incorporating gene expression and environment improves genomic prediction of wheat





Dr Jia Liu, University of Australia and Agriculture and Food, CSIRO | 27 June 2024, Soest

- **Selective breeding** is the process of directed mating to enhance or maintain genetics of desirable traits
- **Wheat** --- Improving *Yield* (height,flowering time, etc)
- **Dominant** components

  to wheat yield?

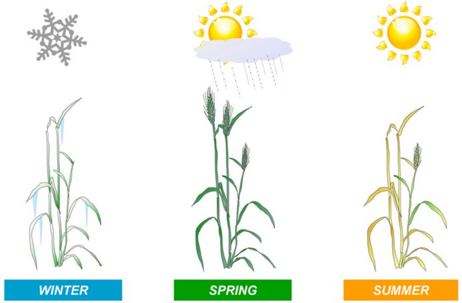
  G, E, GxE

Genomic selection

Genome wide DNA marker data

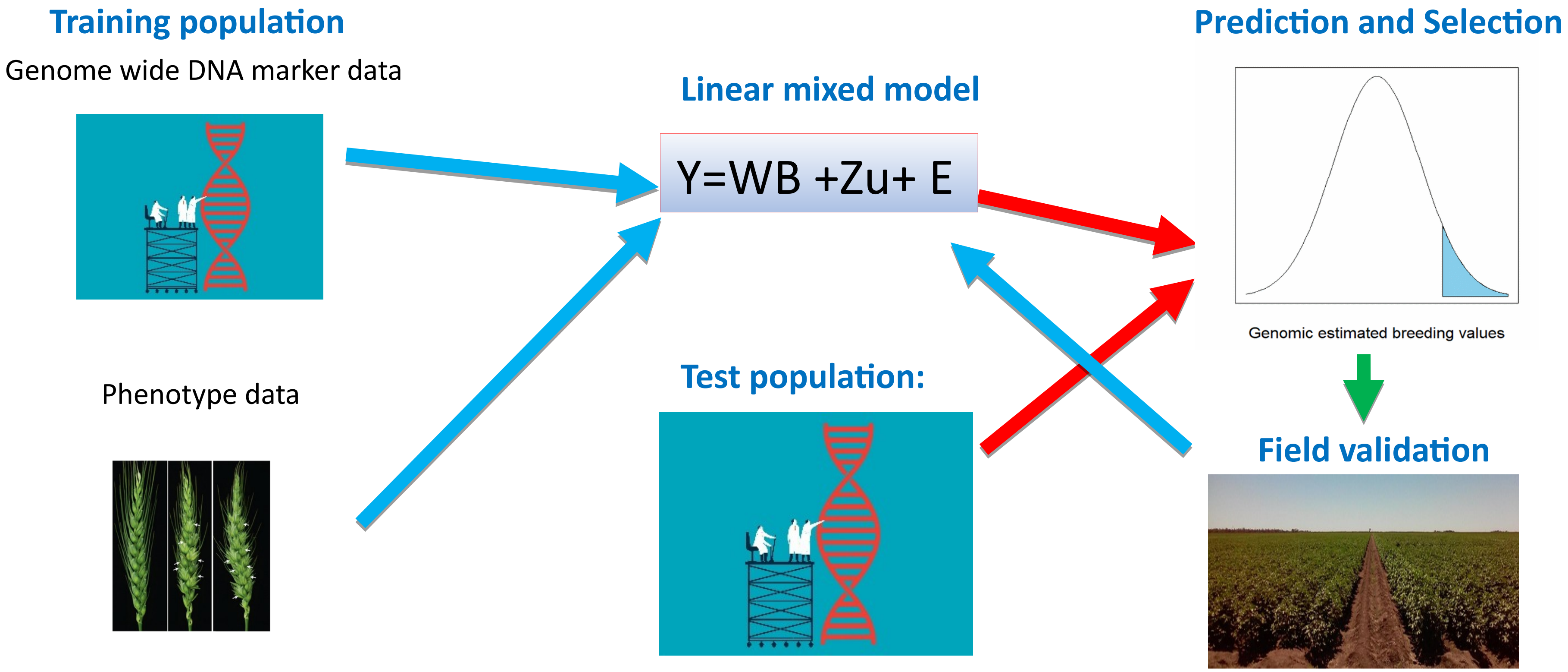
Environment varies

Phenotype data



Field validation

Genomic selection - overview



Training population

Genome wide DNA marker data

Phenotype data

Linear mixed model

$$Y=WB+Zu+E$$

Test population:

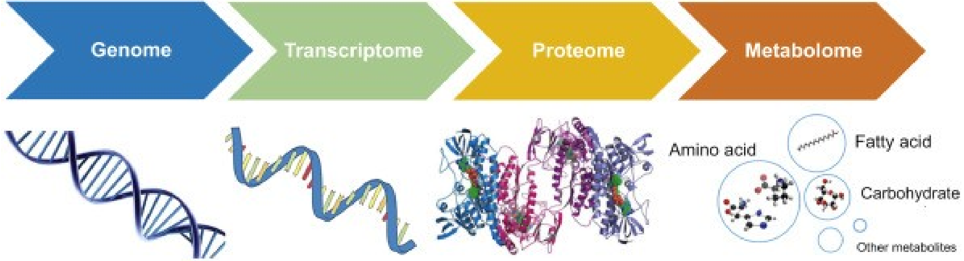
Prediction and Selection

Genomic estimated breeding values

Field validation

- **Gene expression ---phenotypic data**

  High dimensional, quantitative, multilayered biological data e.g. transcriptome, proteome, metabolome



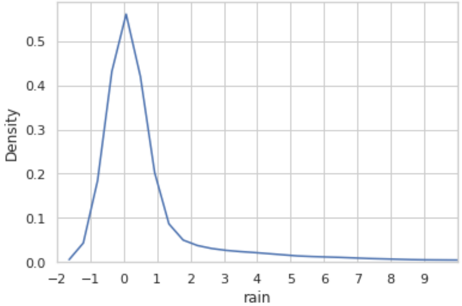
- **Environments**

  More closely linked to the phenotype

- **GxE & high order G**

  Captures other information

  GxE, epistatic effects (high order)



Genemic kinship

- Linear kernel ---GBLUP benchmark predictor

$$K(\boldsymbol{X},\boldsymbol{X}')=\frac{\boldsymbol{X},\boldsymbol{X}'}{trace(\boldsymbol{X},\boldsymbol{X}')/nrow(\boldsymbol{X})}$$

- Nolinear kernel – RKHS

$$k(x_i,x_j)=\exp\left(\frac{\left\|x_i-x_j\right\|^2}{h}\right)$$

$$k(x_i,x_j)=\frac{1}{\pi}|x_i||x_j|\,\sigma(x_i,x_j),$$

- (ML) kernels –neural network kernel (ANN)

activation function $\sigma(x_i,x_j)=\cos^{-1}\left(\frac{x_i x_j}{\|x_i\|\|x_j\|}\right)$

The statistical models:

- LMM

$$\boldsymbol{y} = \mathbf{1}_n \boldsymbol{\mu} + \boldsymbol{Z}_f \boldsymbol{\beta} + \boldsymbol{Z}_g \boldsymbol{u}_g + \boldsymbol{Z}_t \boldsymbol{u}_t + + \boldsymbol{u}_E + \boldsymbol{u}_{gE} + \boldsymbol{e}, where$$

$$\boldsymbol{u}_g \sim N(0, \boldsymbol{K}_g \sigma_g^2), \boldsymbol{u}_E \sim N(0, \boldsymbol{K}_E \sigma_E^2), \boldsymbol{u}_T \sim N(0, \boldsymbol{K}_T \sigma_T^2), \boldsymbol{u}_{gE} \sim N(0, \boldsymbol{Z}_g \boldsymbol{K}_g \boldsymbol{Z}_g' \otimes \boldsymbol{K}_E \sigma_{gE}^2), \boldsymbol{e} \sim N(0, \boldsymbol{I} \sigma_e^2),$$

- Gaussian process $\boldsymbol{y} = \mathbf{1}_n \boldsymbol{\mu} + \boldsymbol{f} + \boldsymbol{e}, \boldsymbol{f}(\boldsymbol{X}) \sim N(\mathbf{0}, \sum_i \boldsymbol{K}_i(\boldsymbol{X}_i, \boldsymbol{X}_i'))$

- Bayesian & frequentist

Bayesian inference

Likelihood

$$p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{\xi})=\prod_{i=1}^{n} N(y_i|f_i,\boldsymbol{\xi}),\ \boldsymbol{\xi}=\{\boldsymbol{\theta},\sigma_e\}.$$

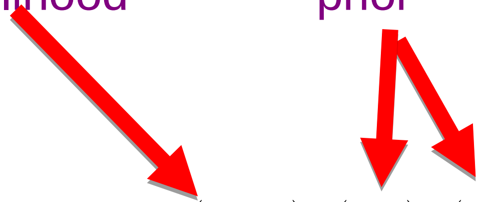
The joint posterior

$$p(\boldsymbol{f},\boldsymbol{\xi}|\boldsymbol{y})\propto\left(\prod_{i=1}^{n} N(y_i|f_i,\boldsymbol{\xi})N(f_i|0,\sigma_u^2 s_i)\right)\prod_{j=1}^{J} p(\xi_j)$$

Marginal Likelihood (ML)

$$p(\boldsymbol{y}|\boldsymbol{X})=\int_f\int_\xi p(y|\boldsymbol{f})\,p(\boldsymbol{f}|\boldsymbol{\xi})\,p(\boldsymbol{\xi})\,d\boldsymbol{f}\,d\boldsymbol{\xi}$$

Bayesian theorem



Likelihood

prior

$$\frac{p\left(\boldsymbol{y}|\boldsymbol{f}\ \xi\right)p\left(\boldsymbol{f}|\xi\right)p\left(\xi\right)}{\int_{f}\int_{\xi}p\left(y|\boldsymbol{f}\right)p\left(\boldsymbol{f}|\xi\right)p\left(\xi\right)d\boldsymbol{f}\,d\,\xi}$$

Marginal likelihood

Test case – Wheat controlled environment

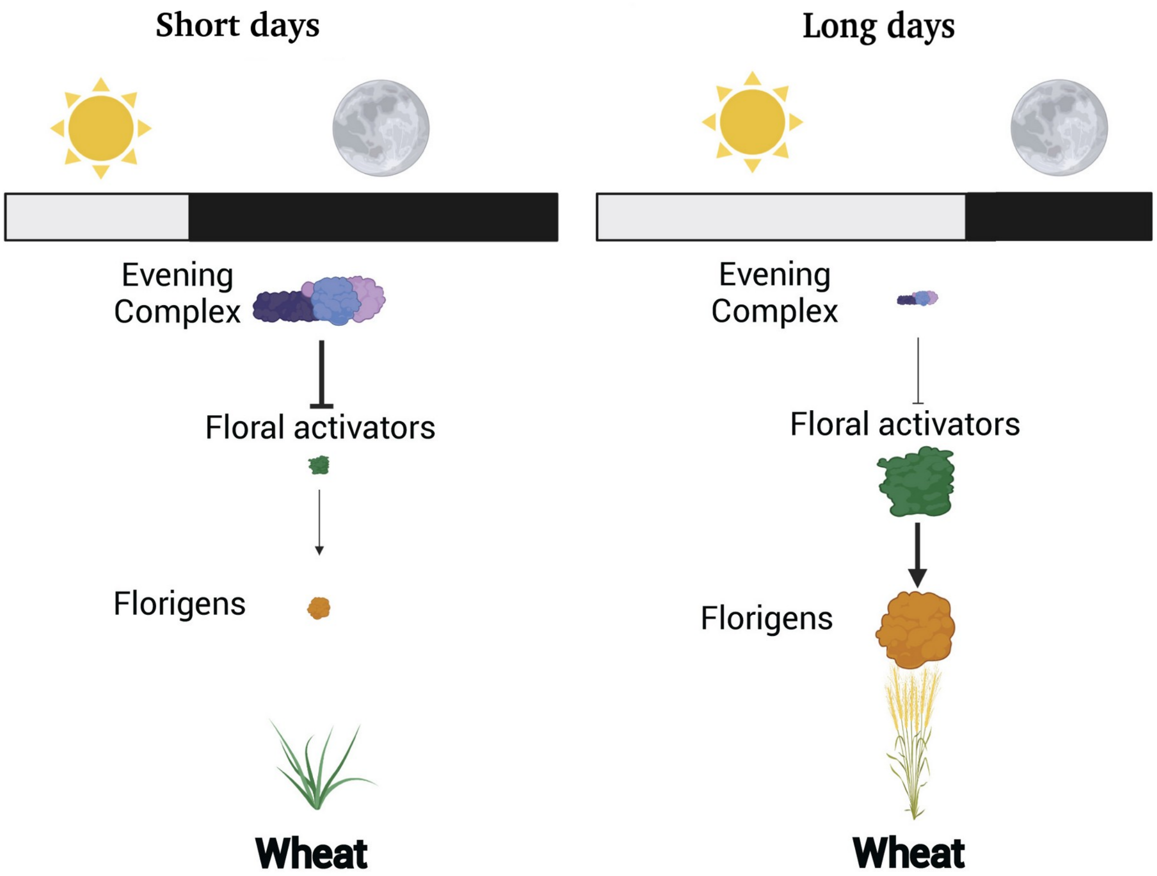
DATA

Transcriptomes generated for 300 varieties in OzWheat
day length - long and short days (16hr/8hr)
600 – transcriptomes
SNPs (~40K)

Genomic prediction models

G-BLUP
Bayesian Gaussian Process
GxE

Long and short days



LMM

Different tested GBLUP models:

1. $y = \mu + G + \varepsilon$
2. $y=\mu + G + GxE + \varepsilon$

genomic effects

3. $y=\mu + T + \varepsilon$
4. $y = \mu + G + T + \varepsilon$
5. $y = \mu + G + T + G\#G + \varepsilon$

+ transcriptome effects

6. $y = \mu + G + G\#G + \varepsilon$
7. $y = \mu + G + A + G\#G + \varepsilon$

+ genomic interactions, epistasis and dominance

8. $y = \mu + G + T + G\#G + GxE + \varepsilon$
9. $y = \mu + G + T + A + G\#G + GxE + \varepsilon$

+ genomic and environmental interaction

“A” shorts for dominant effects,

“G” for genomic effects,

“T” for transcriptome effect,

“E” for environmental effects,

“G#G” is epistasis and “ε” is Gaussian noise.

RKHS

Nonlinear -- Reproducing Kernel Hilbert Space Regression

Four scenarios tested:

1. $y = \mu + G + \varepsilon$

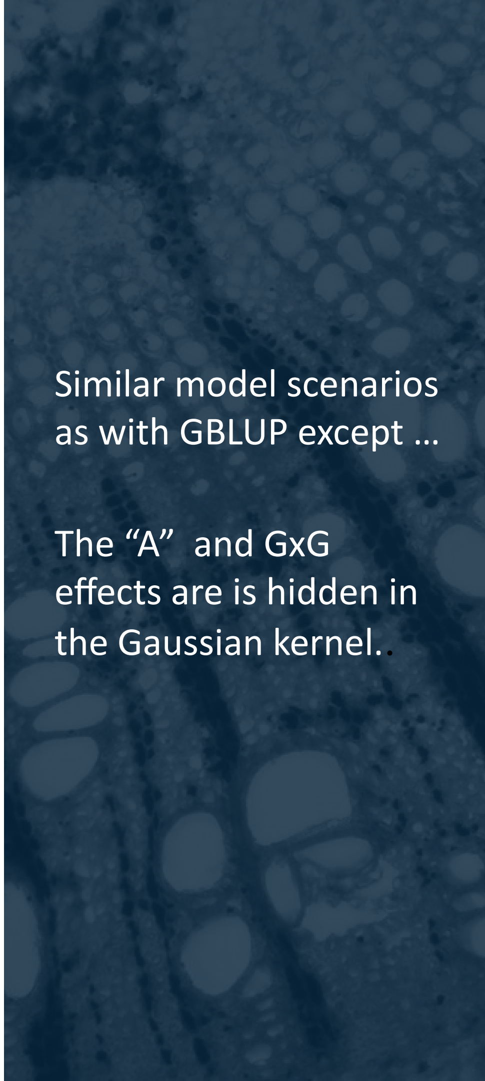
2 . $y = \mu + T + \varepsilon$

genomic effects + A + GxG

3. $y = \mu + G + T + \varepsilon$

4. $y = \mu + G + T + GxE + \varepsilon$

+ genomic and environmental interactions



Similar model scenarios as with GBLUP except …

The "A" and GxG effects are is hidden in the Gaussian kernel..

Model assessment

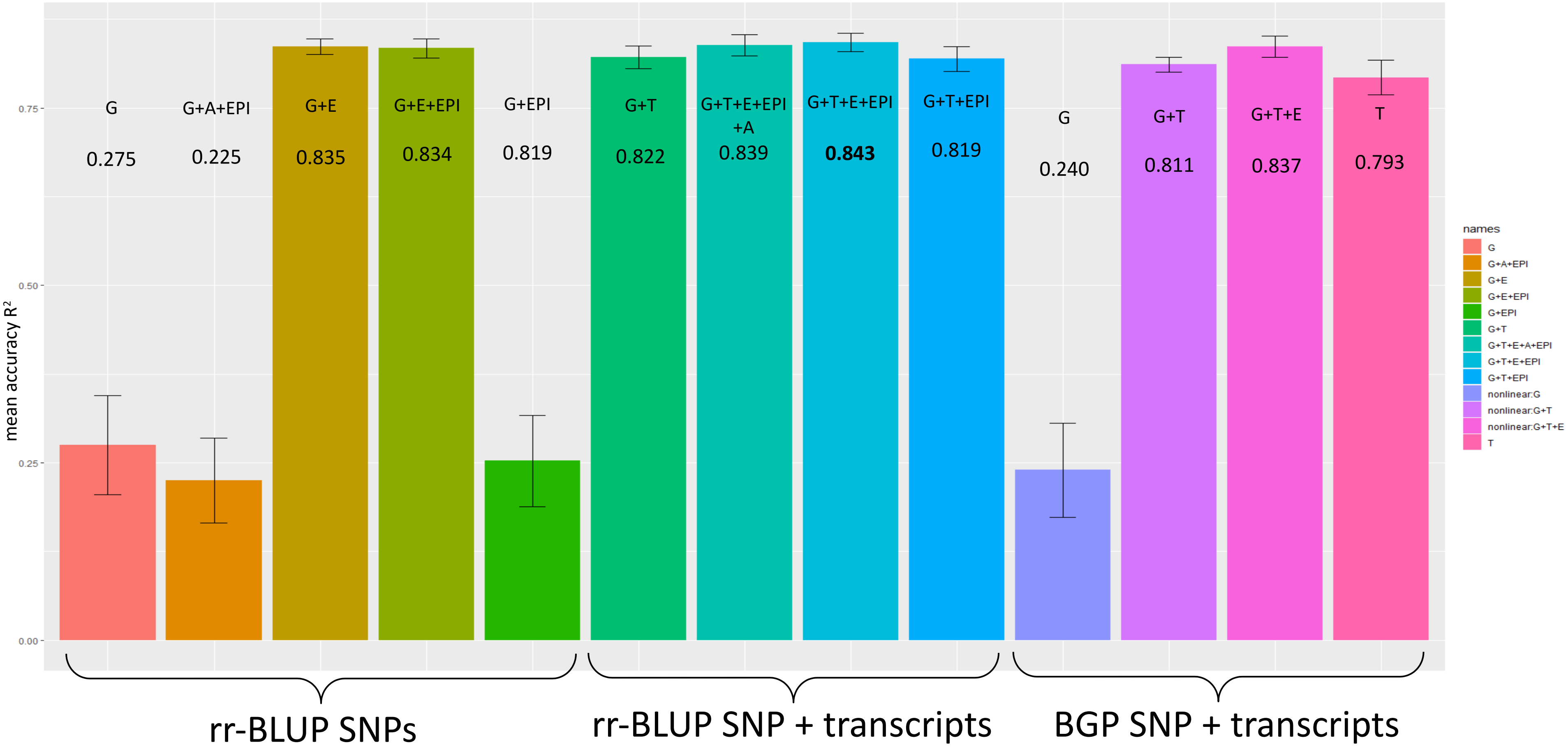
Defining training and test population:

- Cross validation
- Validation set approach
- Simulation – GT
- Historical data / optimizing the object fun/loss

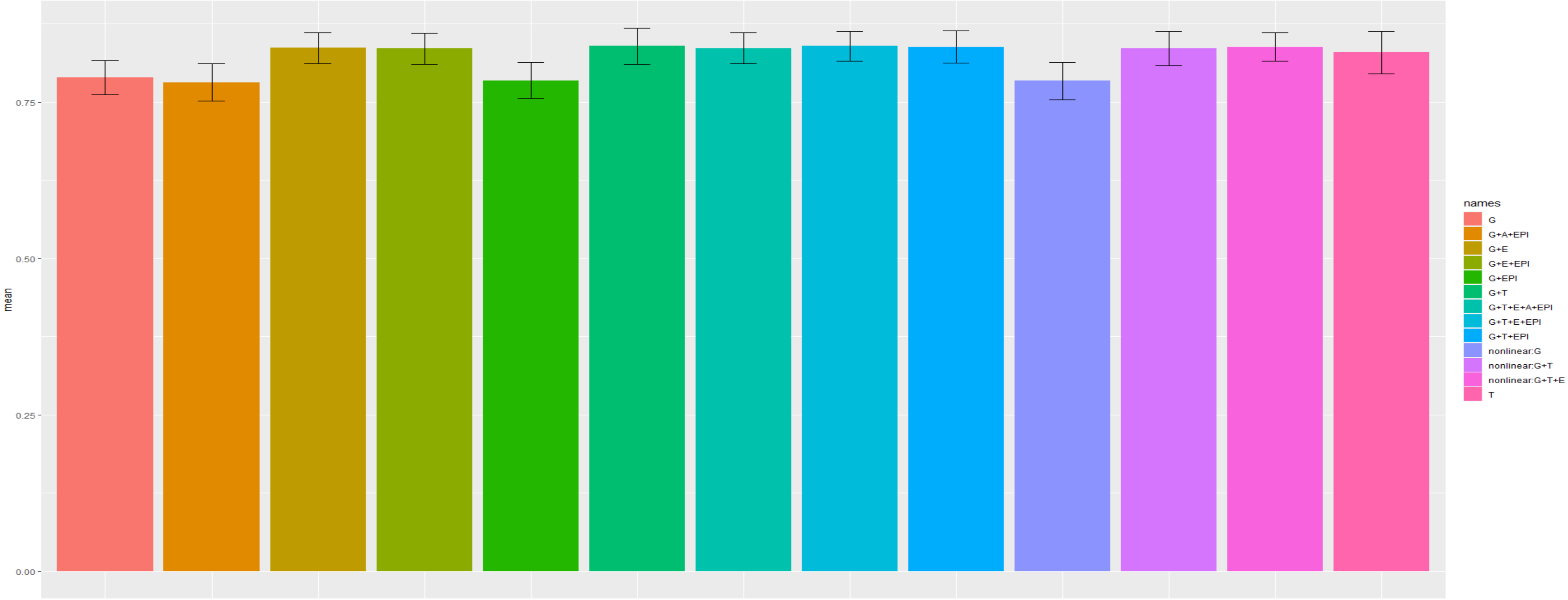
Predictive accuracy:

- Pearson correlation between GEBV and true phenotypes
- RMSE

Flowering time



Height



mean
0.00
0.25
0.50
0.75
names
G
G+A+EPI
G+E
G+E+EPI
G+EPI
G+T
G+T+E+A+EPI
G+T+E+EPI
G+T+EPI
nonlinear:G
nonlinear:G+T
nonlinear:G+T+E
T

	Height	FT
G	0.789 (0.027)	0.275 (0.070)
G+E	0.837 (0.025)	0.837 (0.011)
G+EPI	0.785 (0.029)	0.253 (0.064)
G+E+EPI	0.836 (0.024)	0.834 (0.013)
G+A+EPI	0.782 (0.030)	0.225 (0.060)
T	0.830 (0.034)	0.793 (0.024)
G+T	**0.840 (0.030)**	0.822 (0.016)
G+T+EPI	0.838 (0.026)	0.819 (0.017)
G+T+E+EPI	**0.840 (0.023)**	**0.843 (0.013)**
G+T+E+A+EPI	0.836 (0.025)	0.839 (0.015)
nonlinear:G	0.784 (0.030)	0.240 (0.067)
nonlinear:G+T	0.836 (0.028)	0.811 (0.011)
nonlinear:G+T+E	0.838 (0.023)	0.837 (0.015)

Conclusion

The best model included all effects

- genome SNP, transcriptome, GxE and G#G
- including transcriptome improved model performance

Transcriptome approximated GxE

- categorical environments did not play a critical role in the prediction when transcript data were included
  - e.g. good performance of G+T and T only models (why)
- but did not outperform the combined model, suggesting explicit characterisation of GxE and GxG is warranted
- may be useful ---E (No records, poorly characterized/collected)
  - otherwise G+E sufficient

Nonlinear and linear kernel perform similarly here

- data of small size/scale
- environment covariates are simple
- Plaint kernel structure

From: gaussian-process-not-quite-for-dummies

Software

R package

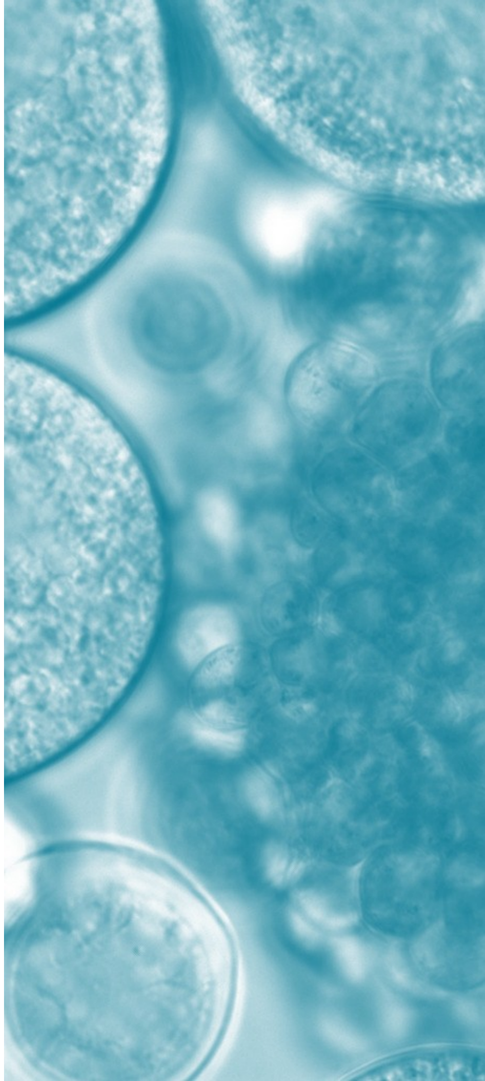
- rrBLUP, BGLR, BGGE

Discussions

- Deep learning to work with large-scaled genomic SNP data, e.g. graphic neural network, deep Gaussain process

- Fast computational algorithms to handle large densed kernel matrices; convolution transformation to ease the computational burden.

- Kernel development & selection



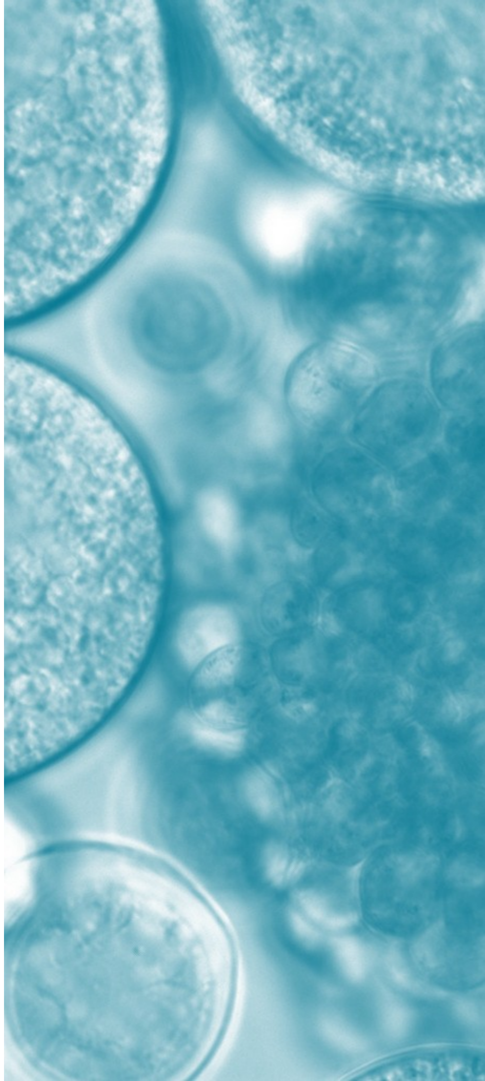
Reference

- Z. Li, N. Gao, J. W. R. Martini, H. Simianer, Integrating Gene Expression Data Into Genomic Prediction, Frontiers in Genetics 10 (2019).

- Costa-Neto, G., Fritsche-Neto, R., & Crossa, J. (2021). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity, 126(1), 92–106. https://doi.org/10.1038/s41437-020-00353-1

- D. J. Tolhurst, R. C. Gaynor, B. W. Gardunia, J. M. Hickey, G. Gorjanc, Genomic selection using random regressions on known and latent environmental covariates, The oretical and Applied Genetics 135 (2022) 3393–3415

- Gaussian Processes for Machine Learning [book], CE Rasmussen, C KI Williams (2006), MIT



Thank you very much!