Australasian Applied Statistics Conference 2024, Rottnest Island, Western Australia

An efficient Bayesian dimensional reduction regression method for multiple environment genomic prediction

Zitong Li | 04/09/2024

Australia's National Science Agency

Structure of this talk

- 1. Introduction to the application problem:

-Cotton breeding in Australia

-Genomic selection

- 2. Statistical solution:

-Bayesian high dimensional linear regression

- 3. Results; Conclusion

Background: Australian cotton production

Highest yield

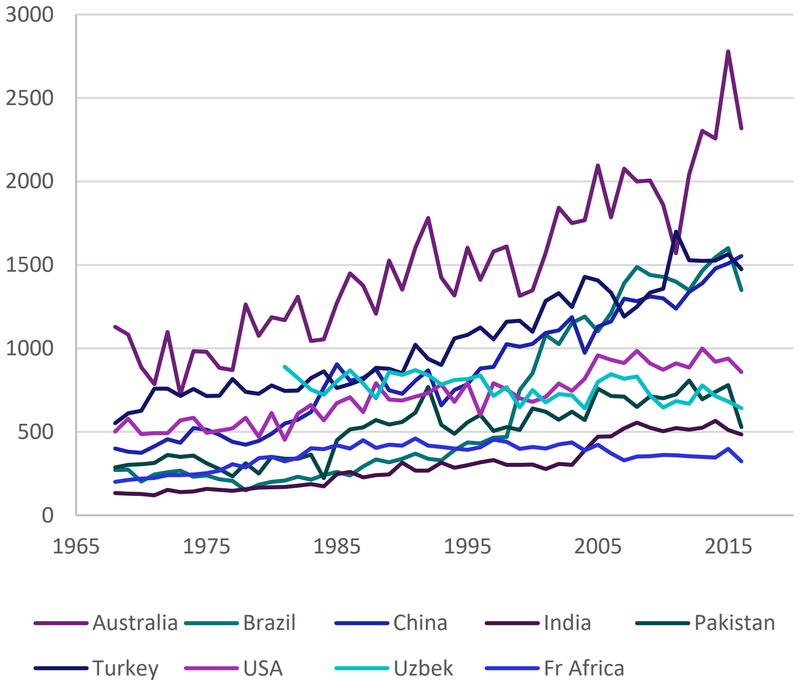
Australian cotton yield is the highest, 3 times the global average

Highest quality

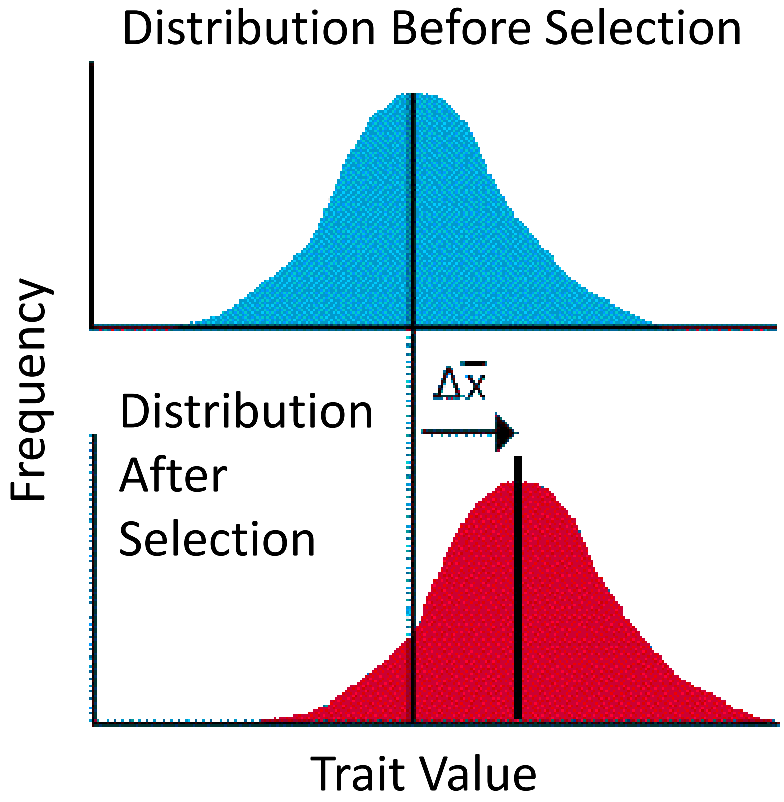
Australian cotton has a global reputation for high quality, >40% above base grade

CSIRO’s contribution

Developing 100% of Australian cotton varieties through our cotton breeding program

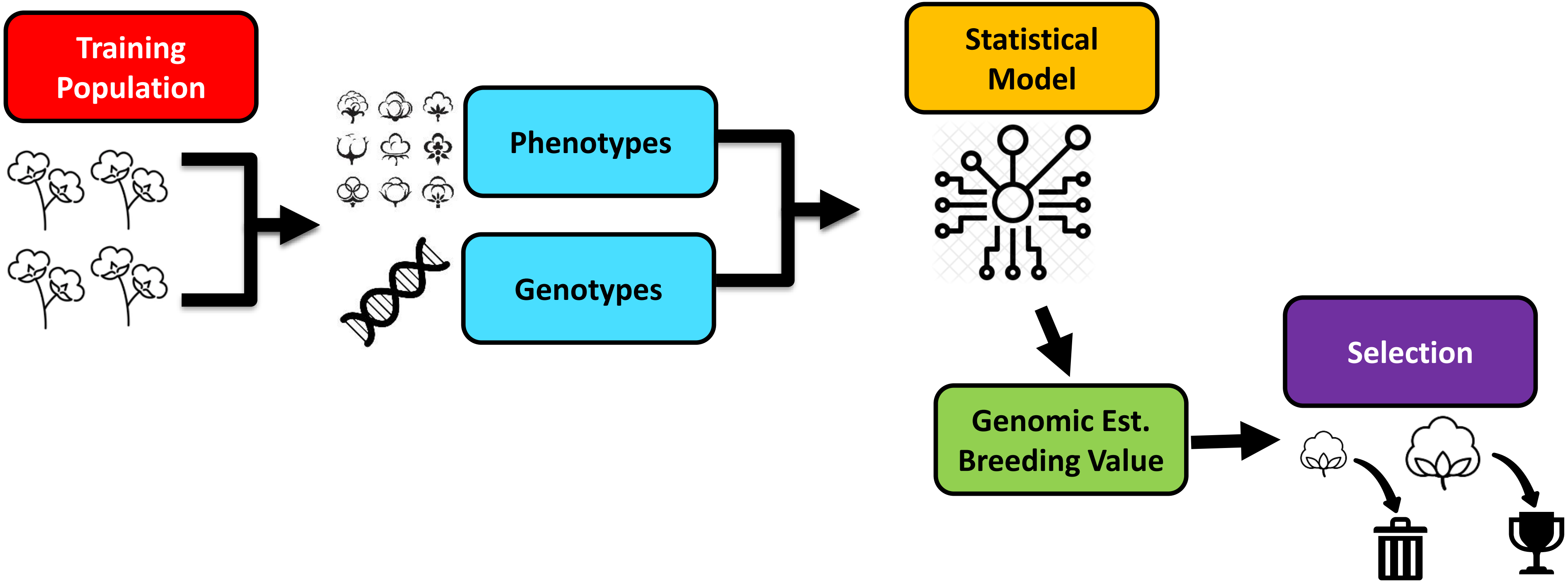


Conventional cotton Breeding



- Conventional approach
  - Phenotype selection
  - Disadvantage: breeding cycle is long
  - Usually over 10 years to introduce a new cotton variety
  - Use DNA information instead?
  - If DNA info can accurately predict phenotypes, we can select variety at early stage of plant development

Genomic Selection



Training Population

Phenotypes

Genotypes

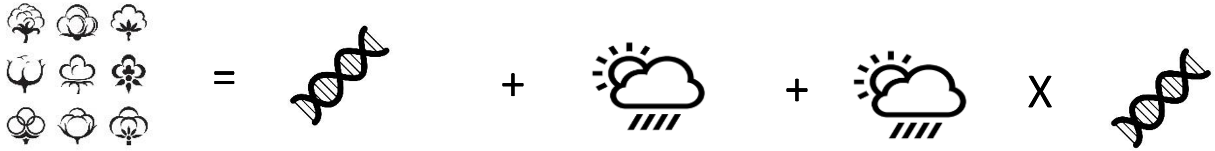
Statistical Model

Genomic Est. Breeding Value

Selection

Challenges

- In plant breeding, phenotype variation is explained by both **genetics**, **environment**, and possible **interactions** between them. So it will be crucial to account for all these factors in any quantitative genetic model for plant breeding.



- Dimensionality became ultra high when considering high resolution genomic and environmental data, especially the interaction between them.

-e.g. 100 00 genetic markers × 50 environmental covariates= 500 000 G × E interactions

Data collection

-All data from Australia conventional cotton grow area (New South Wales, Queensland, Victoria).

-9k genetic markers of over 4000 cotton lines. Genotype data distributed over 26 chromosomes

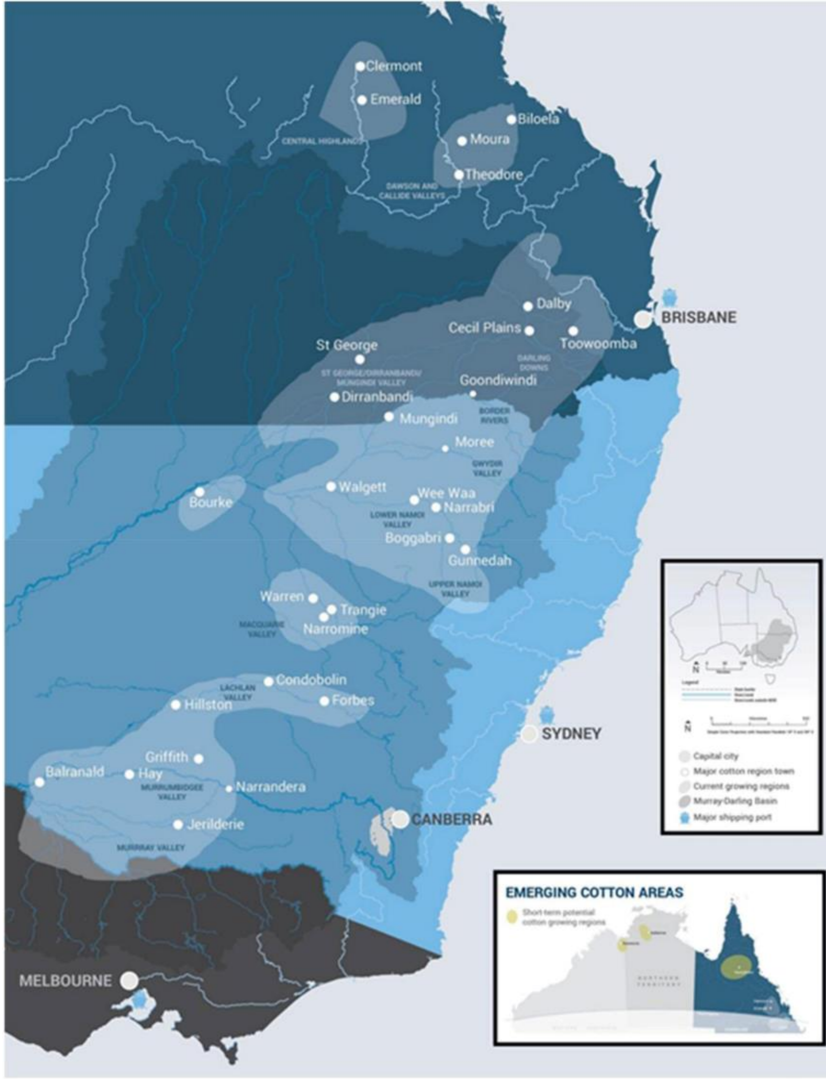
-Phenotyping over multiple years (2012-2022)

-> 77 location-year combinations

-Over 12000 phenotype records

-Lint Yield, Fibre quality traits

-On site weather station to collect climate data: 50+ environmental covariates



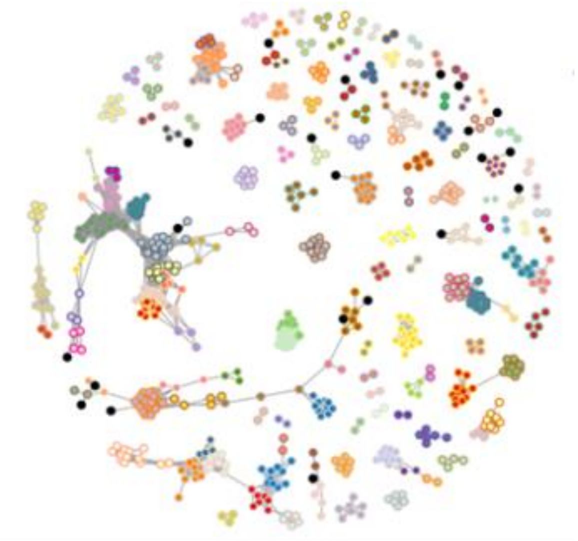
Dimensional reduction using LD clustering

- Linkage disequilibrium (LD) is a phenomenon in genetics: genetic markers at nearby genome locations tend to be more correlated to each other.

- Apply a LD network clustering algorithm to classify genetic markers into groups. And then apply dimensional reduction on genetic data (Li et al. 2018; Molecular Ecology Resources)

- Significantly reduce number of G×E parameters (e.g. 500 000->500 00)

Chromosome



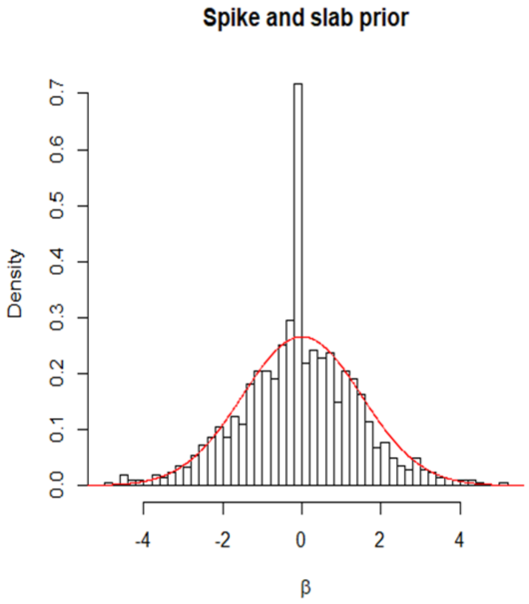
Linkage disequilibrium clustering

Method:

- High dimensional Bayesian regression methods

$$Y = \sum X_E \beta_E + \sum X_G \beta_G + \sum\sum X_E X_G \beta_{G\times E} + e,$$

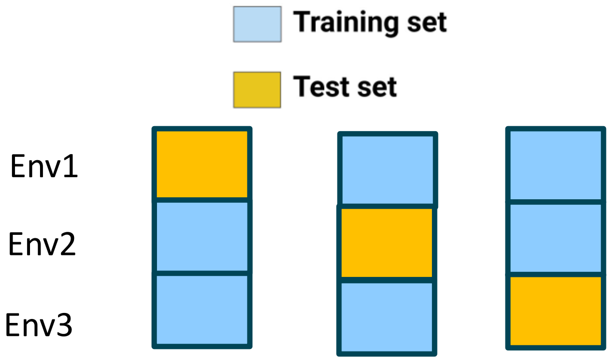
-Use MCMC algorithm to search through the model space. In each round, identify a subset of important variables. Then average over the MCMC samples to get estimation of GEBVs



-$P(\beta_j|\gamma_j) \propto (1-\gamma_j)I_{(\beta_j=0)} + \gamma_j N(\beta_j|0,\sigma_j^2)$,

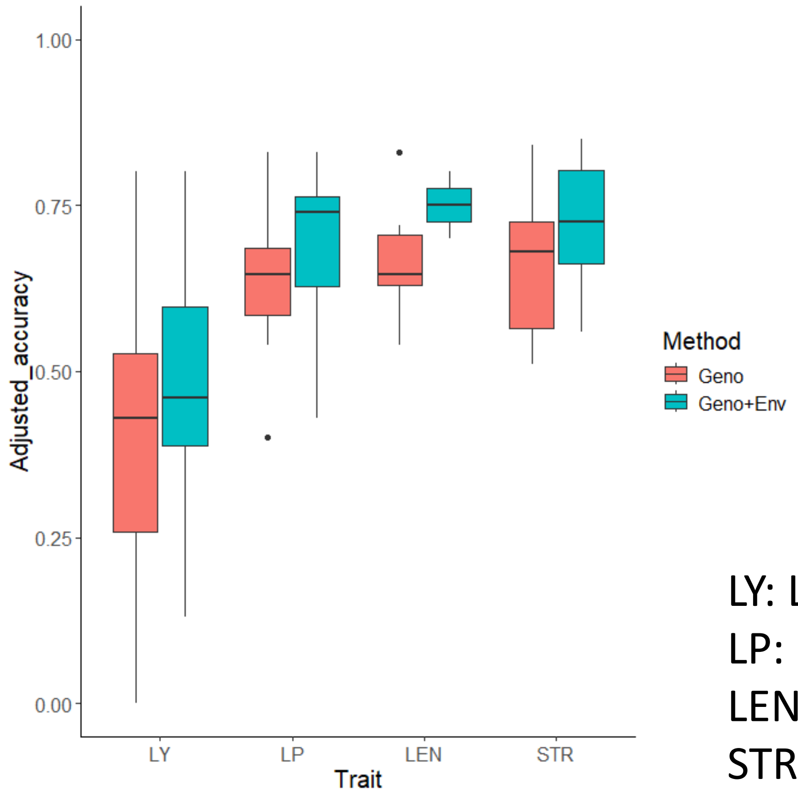
-$p(\gamma_j|\omega) = \omega^{\gamma_j}(1-\omega)^{1-\gamma_j}$

Genomic prediction results



Leave one environment out analysis

Prediction accuracy: $\dfrac{\text{Cor(GEBV,true phenotype)}}{\text{square root(heritability)}}$



LY: Lint yield

LP: Lint percentage

LEN: Fibre length

STR: Fibre strength

Summary

- The genomic prediction shows potential to accurately predict both yield and fibre quality traits.
- Proposed Bayesian model can efficiently analyse our data sets using 4-5 hours time.
- More interpretable compared to other methods such as G-BLUP or FA models.
- Also investigating on other methods such as deep neural network.
- GS starts to be deployed in our breeding program.

- **References:**

-Li et al. (2024) Manuscript under preparation

-Rafter et al. (2024) under revision in Field Crops Research

-Khalilisamani and Li et al. (2024), Frontiers in Plant Science

-Li et al. (2024) Theoretical and Applied Genetics

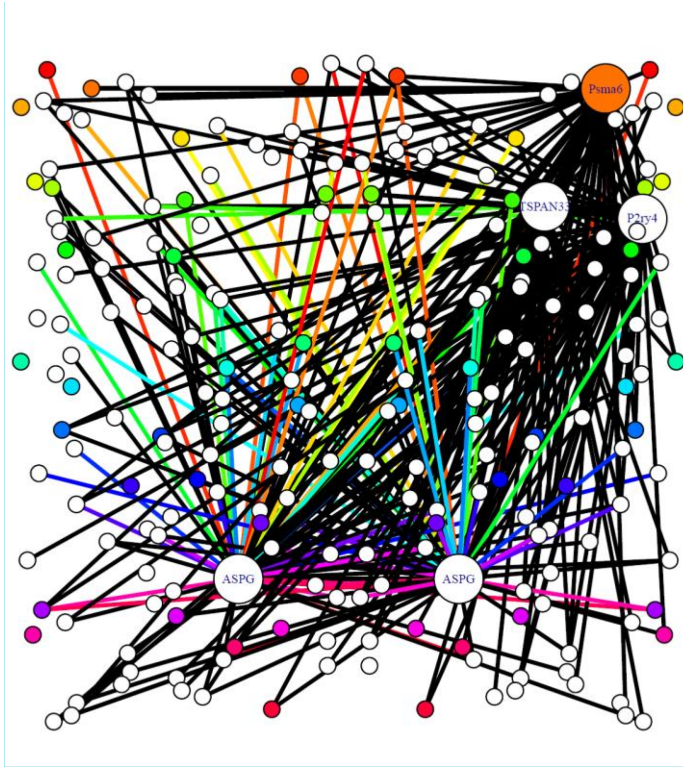
-Li and Gutierrez (2023) Frontiers in Genetics.

-Li et al. (2022) Heredity.

Also See Dr Ngoc Dung Nguyen's poster (next to coffee table)!

- Gaussian graphical models to construct gene expression networks on the basis of heterogenous population

- Manuscript under revision in Asian Conference on Machine Learning (ACML)



Thank you

- CSIRO Agriculture & food research unit
  - Zitong Li
    Research Scientist

  +61 2 6246 5235

  zitong.li@csiro.au

- **Acknowledgement:**

- Shiming Liu
- Warren Conaty
- Warwick Stiller
- Iain Wilson
- Ngoc Dung Nguyen
- Qian-Hao Zhu
- Xuesong Li
- Pierce Rafter
- Philippe Moncuquet
- Angel Popa-Baez