Trait genetic architecture and population structure determine model selection for genomic prediction in natural *Arabidposis thaliana* populations

Patrick Gibbs, Jeff Paril, Alex Fouriner-Level

THE UNIVERSITY OF MELBOURNE

Background

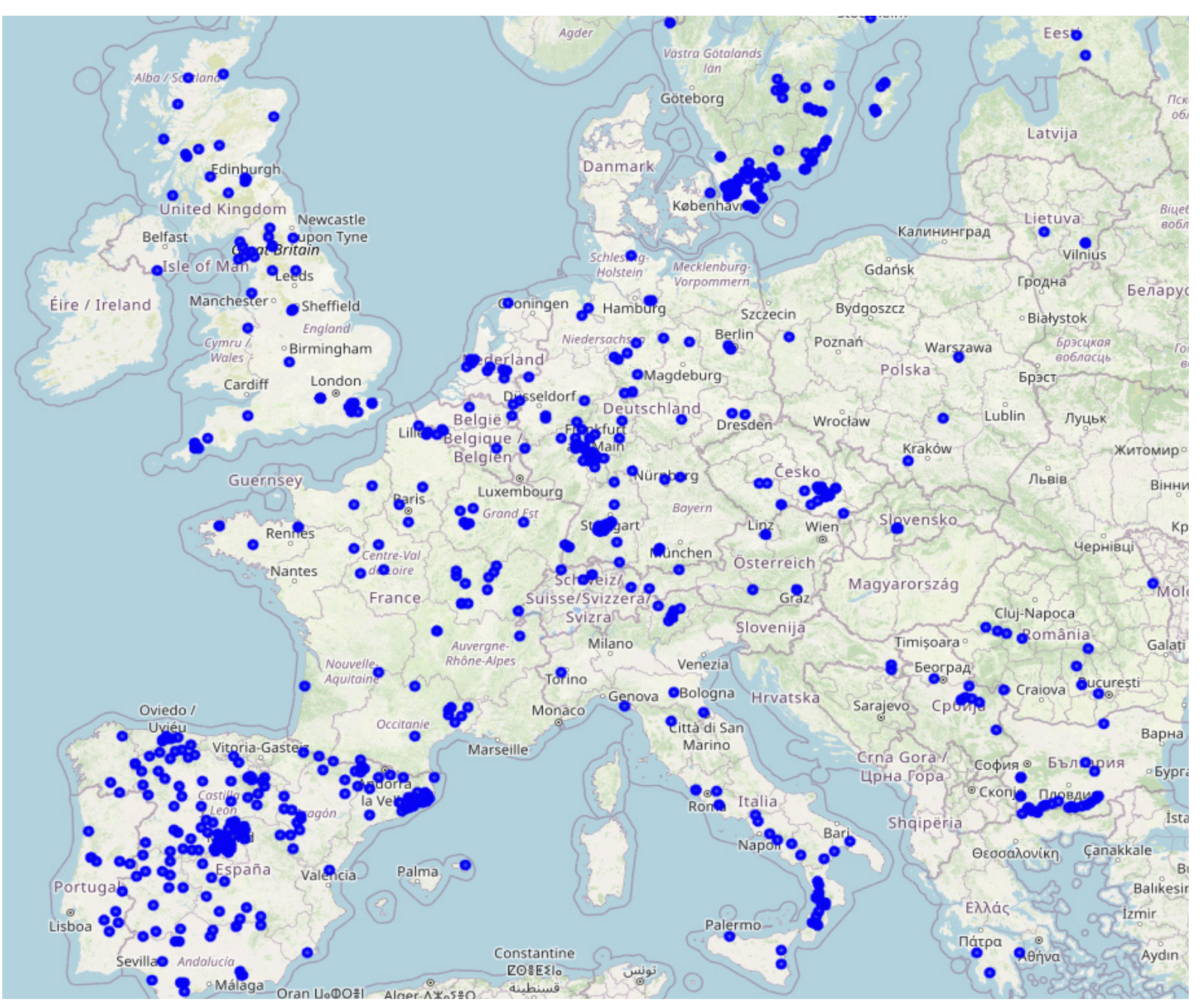
In plants, genomic prediction is used to predict agronomically relevant traits from DNA markers. Agronomically relevant traits vary in genetic architecture, and genomic prediction models are fitted to training populations designed to maximize diversity. Therefore, critical in GP is choosing the most appropriate model for a trait's distribution of genetic effects and the population's allele frequencies.

Aims

1. Understand the relationship between model choice and genetic architecture in genomic prediction.
2. Understand how population structure and demography affects genomic prediction model choice.

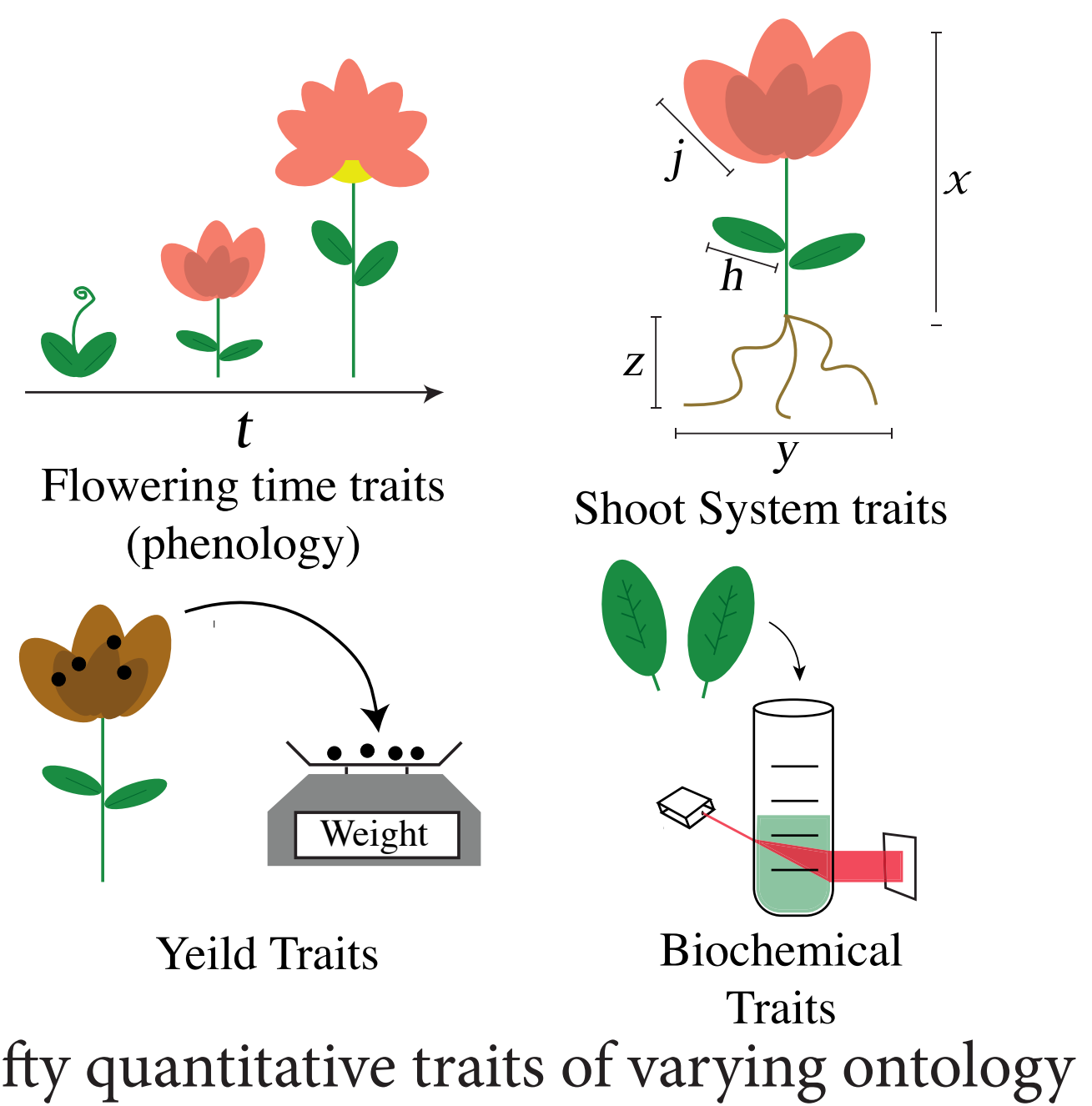
Methodology

1. Sample 1058 *Arabidopsis thaliana* accessions

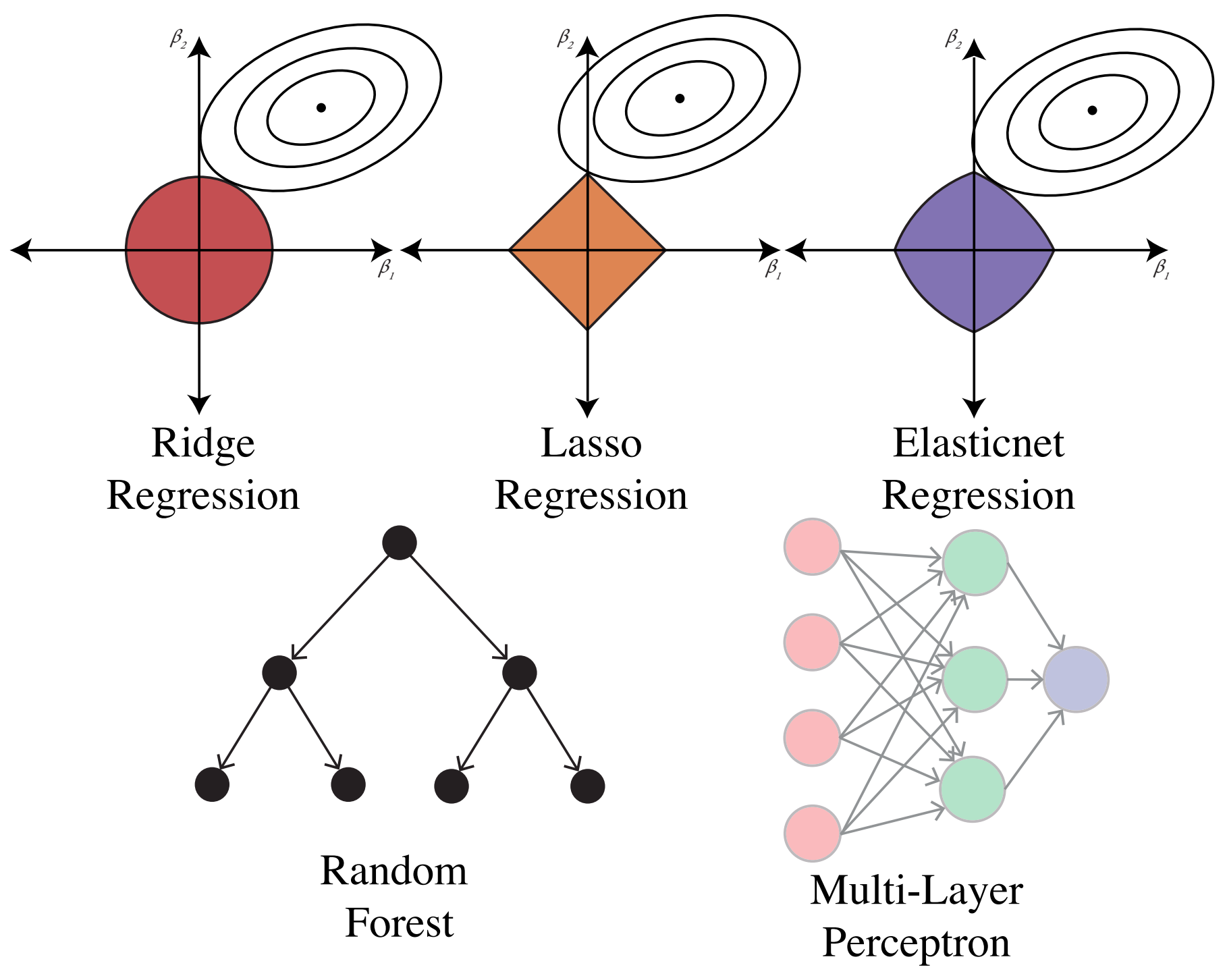


(10001 Genomes Dataset)

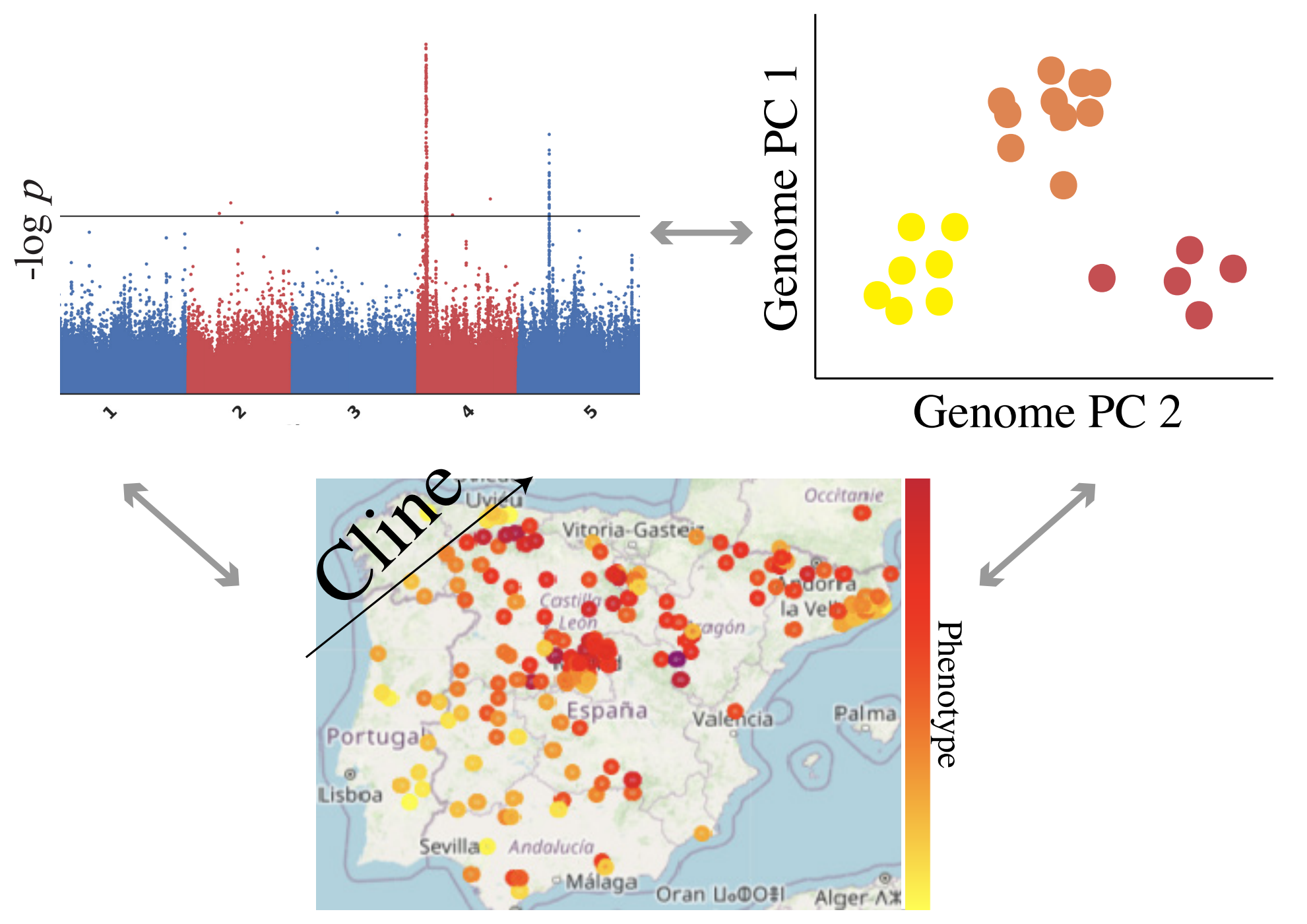
2. Measure sensitivity to trait ontology and genomic prediction model



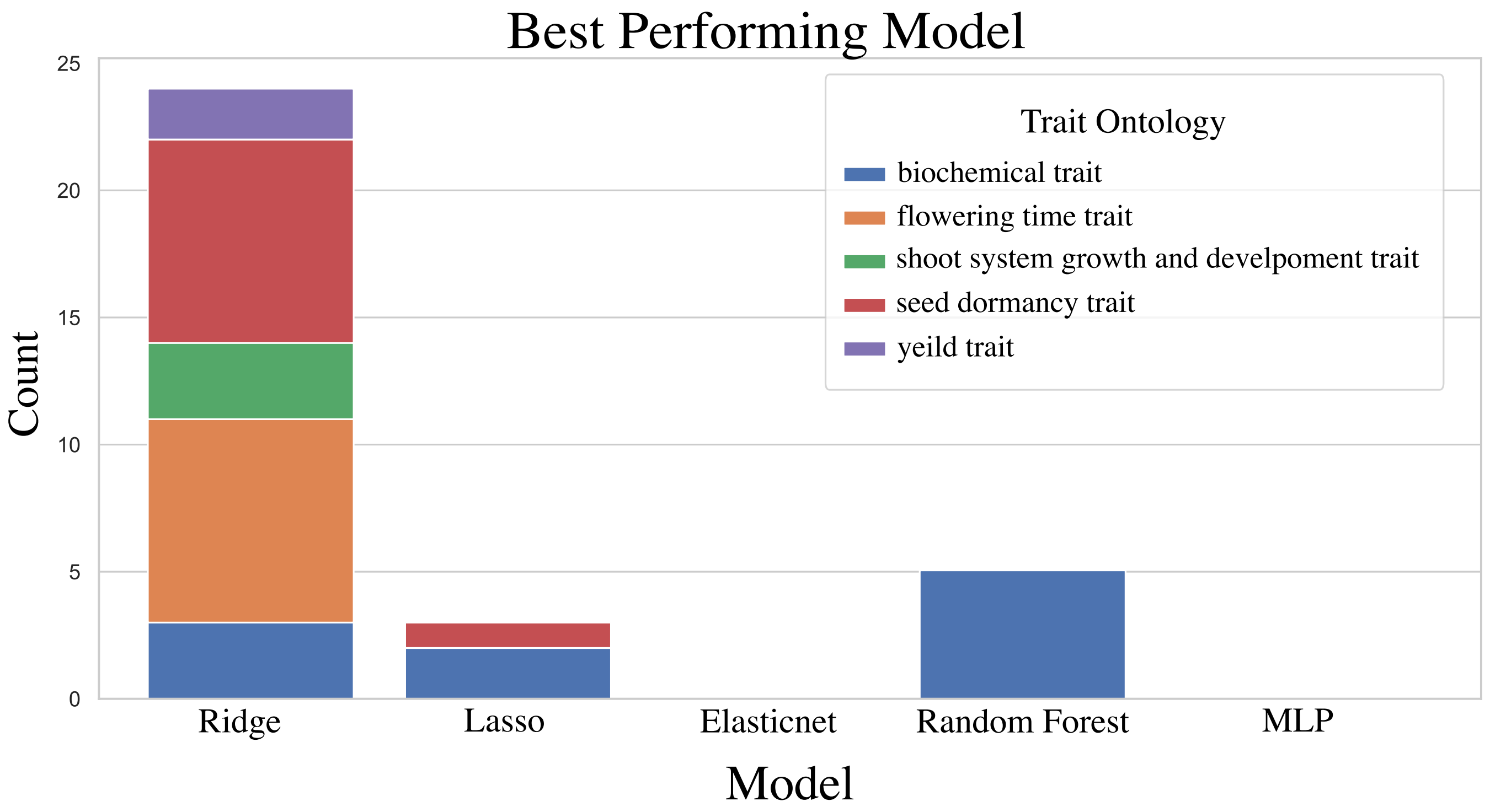
Fifty quantitative traits of varying ontology
(Arapheno Database)



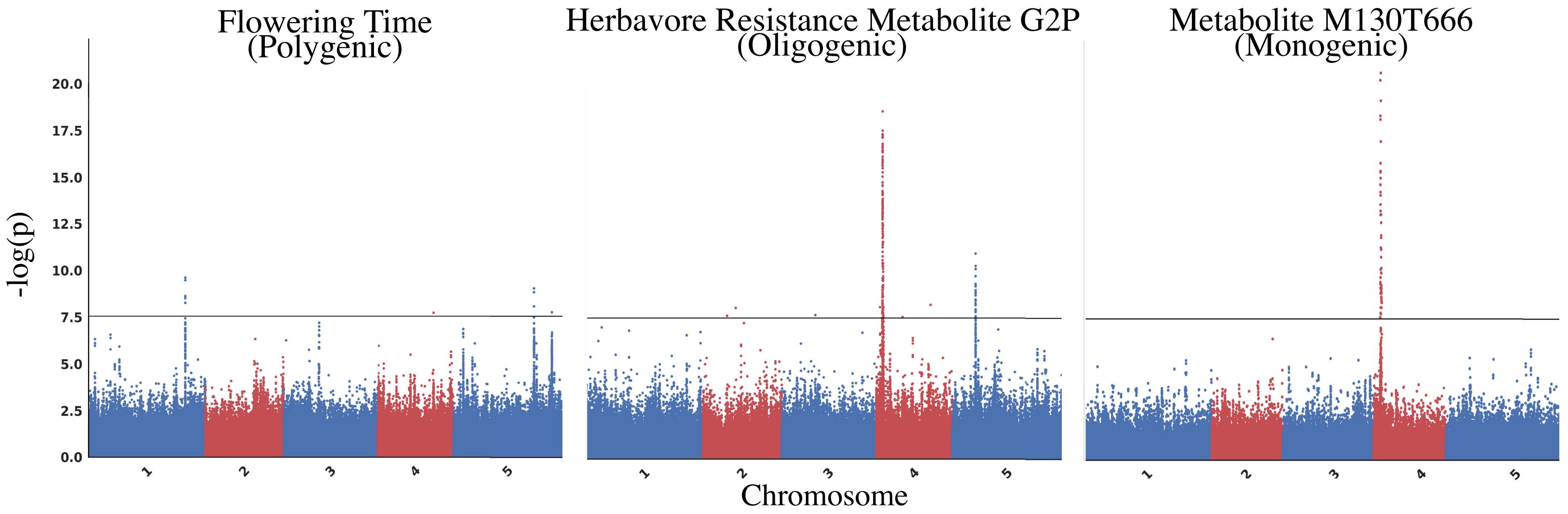
3. Understand the effect of genetic architecture, population structure and selection in genomic prediction experiments



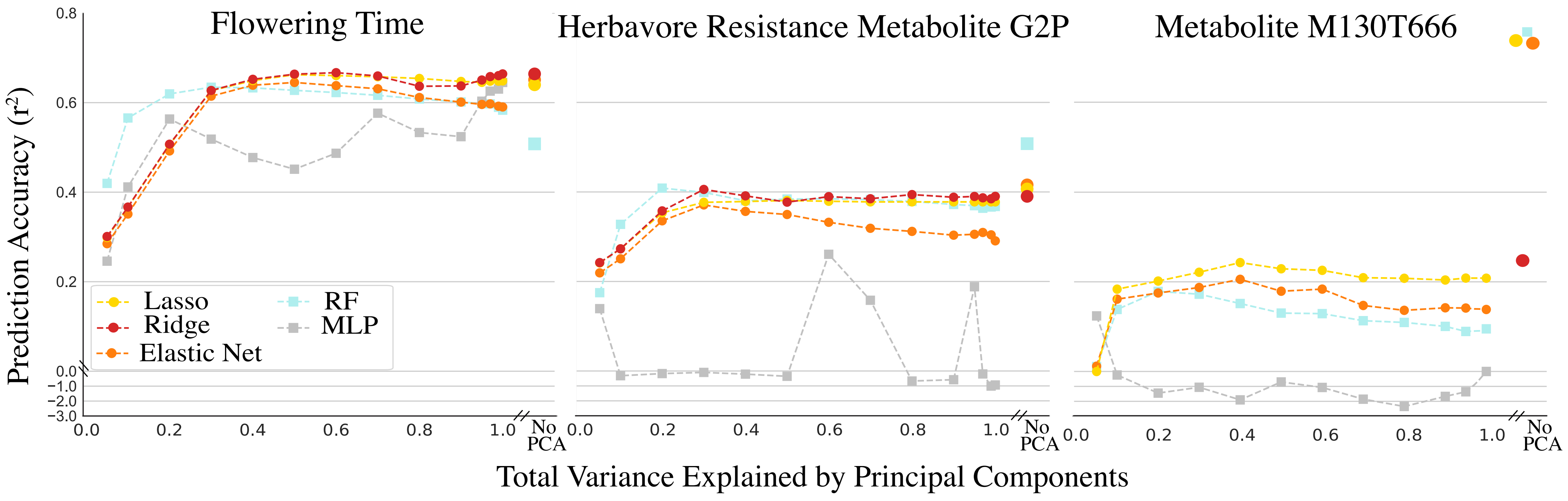
Results



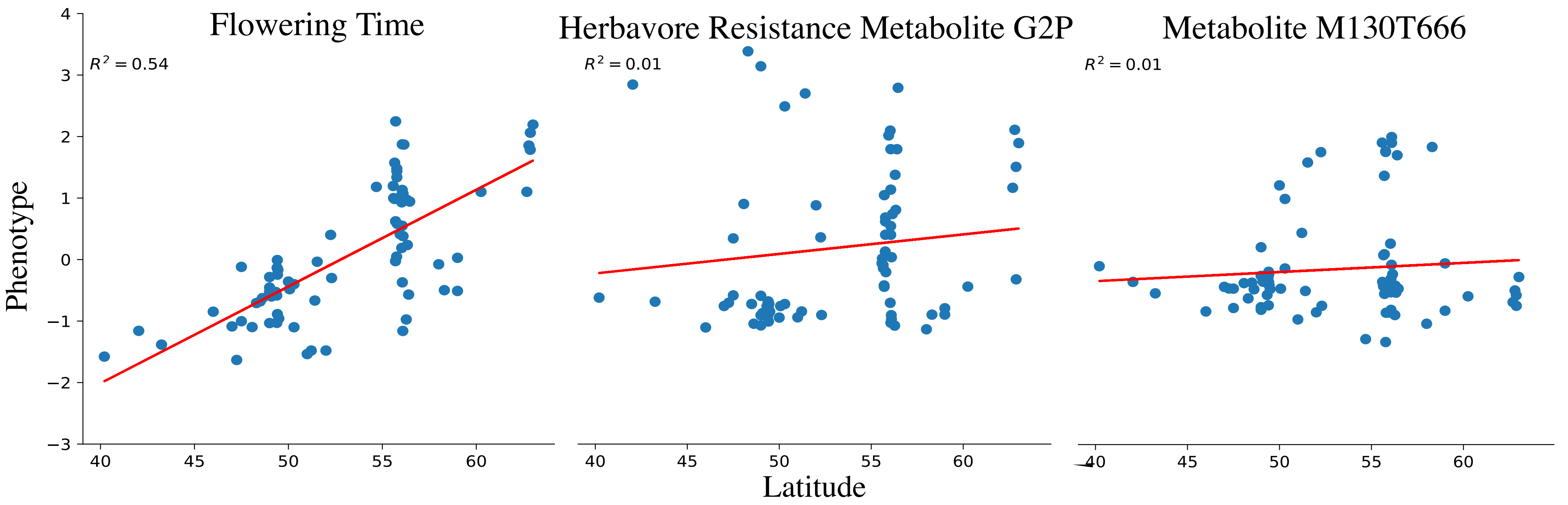
1. Linear statistical models best predicted most traits. However, random forest better predicted some biochemical traits.



2. This can be explained by the fact that biochemical traits tended to have a simpler genetic architecture / are strongly associated with a small number of genomic regions.



3. Contrastingly, macroscopic traits (e.g. flowering time) tended to be polygenic and could be well predicted by population structure in simple linear models. Resolution of individual markers was not required to predict these traits accurately.



4. Many complex (macroscopic) traits covaried with geography, which also explains why these traits are particularly associated with population structure.

Take aways

1. Machine learning approaches like random forest normally fall short in genomic prediction, while linear models dominate. However, in this study we show ensemble models have utility in some simpler (typically metabolic) plant traits.

2. Complex traits (particularly flowering time) covary with both population structure and geography, negating the importance of individual markers in prediction. However, because microscopic traits are less polygenic and covary less with population structure, the state of individual markers is important to predict them.

References

Gentyped accessions: Arouisse, B., Korte, A., van Eeuwijk, F., & Kruijer, W. (2020). Imputation of 3 million SNPs in the Arabidopsis regional mapping population. The Plant Journal: For Cell and Molecular Biology, 102(4), 872–882. https://doi.org/10.1111/tpj.14659

Phenotypic traits sourced from: Seren, U., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K., & Korte, A. (2017). AraPheno: A public database for Arabidopsis thaliana phenotypes. Nucleic Acids Research, 45(D1), D1054– D1059. https://doi.org/10.1093/nar/gkw986