Improving the efficiency of QTL/GWAS analytics

Beata Sznajder[1], Julian Taylor[1]

[1]Analytics for the Australian Grain Industry, Adelaide University

Background

Genome-wide association study (GWAS) methods aim to identify Quantitative Trait Loci (QTL) that underpin phenotypic traits. It enables crop breeding programs to target genomic regions responsible for improved yield or genetic resistance and tolerance to biotics and abiotic stressors (disease, drought, heat or frost). However, GWAS of dense genetic arrays typed across large germplasms entails challenges due to computational efficiency and a good control of type I and II errors.

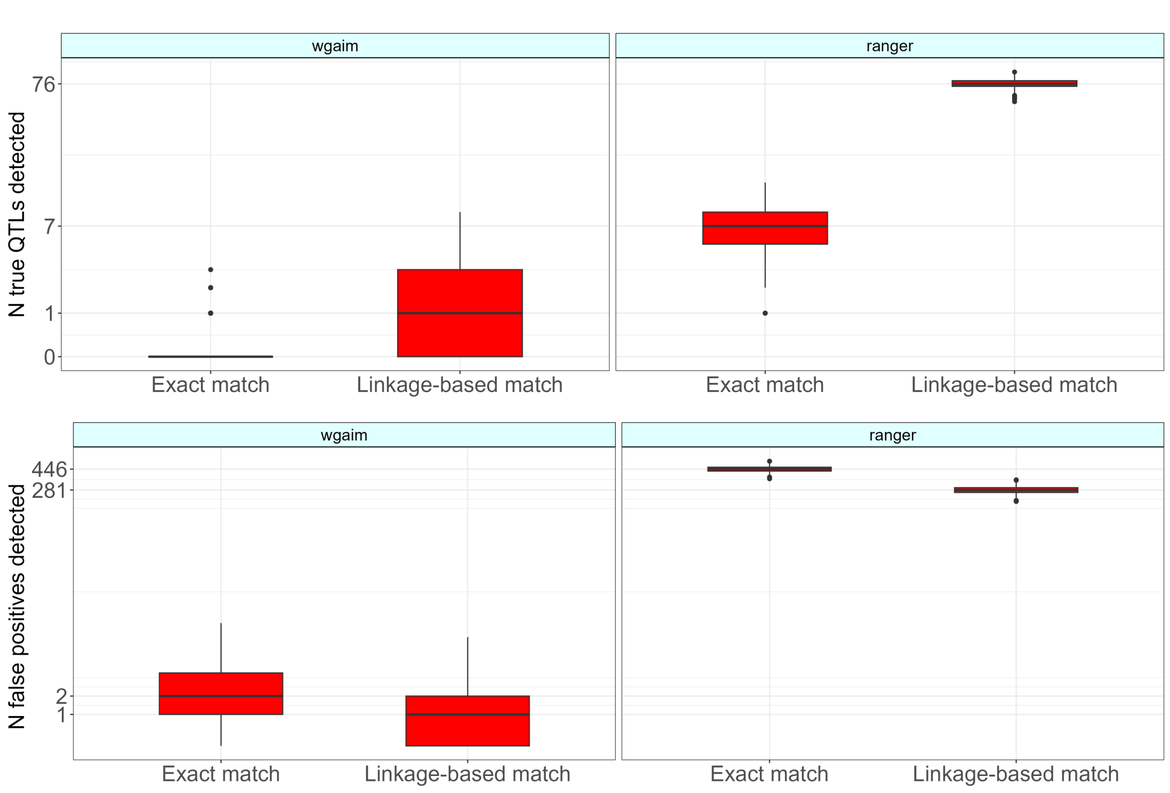
Methodological advances, such as Machine Learning (ML), offer opportunity to depart from single-marker associations towards multi-locus approaches that target both additive and epistatic genetic effects while retaining computational efficiency, biological interpretability and a control of error rate. Here we present our research on improving the efficiency of GWAS analysis for plant breeding, in two-fold approach presented below.

Recent GWAS research

- We developed a **simulation framework** for comparative study of GWAS that uses synthetic plant genetic and phenotypic data simulated with packages `alphaSimR` and `fieldSimR`.
- We conducted comparative GWAS using linear mixed-model based approach implemented by package `wgaim` against ML method Random Forest (RF) implemented by package `ranger`, under low signal-to-noise ratio motivated by real Australian trial data, with key parameters of 200 genotypes, heritability $h^2 = 0.1$ and a 100 true additive QTLs among 100,000 simulated genetic markers.

QTL detection by RF is offset by a very high false positive rate

wgaim provides a balance of statistical power and type I error rate in small-to-medium studies with low signal-to-noise ratio



`ranger` parameters were mtry=250, num.trees=1000, importance="permutation". For calculation of variable importance num.permutations was set to 100.

In linkage-based match, a detected QTL was classified as true QTL if it resided within a region of a true QTL defined as true QTL position ± a window. The window size was based on linkage disequilibrium decay and set to distance at which marker pairwise correlation coefficients reached value 0.2.

y-axis is plotted on logarithmic scale ($\log_2(y+1)$) for more compact display, with the axis labels showing the original untransformed values.

Developing GWAS research

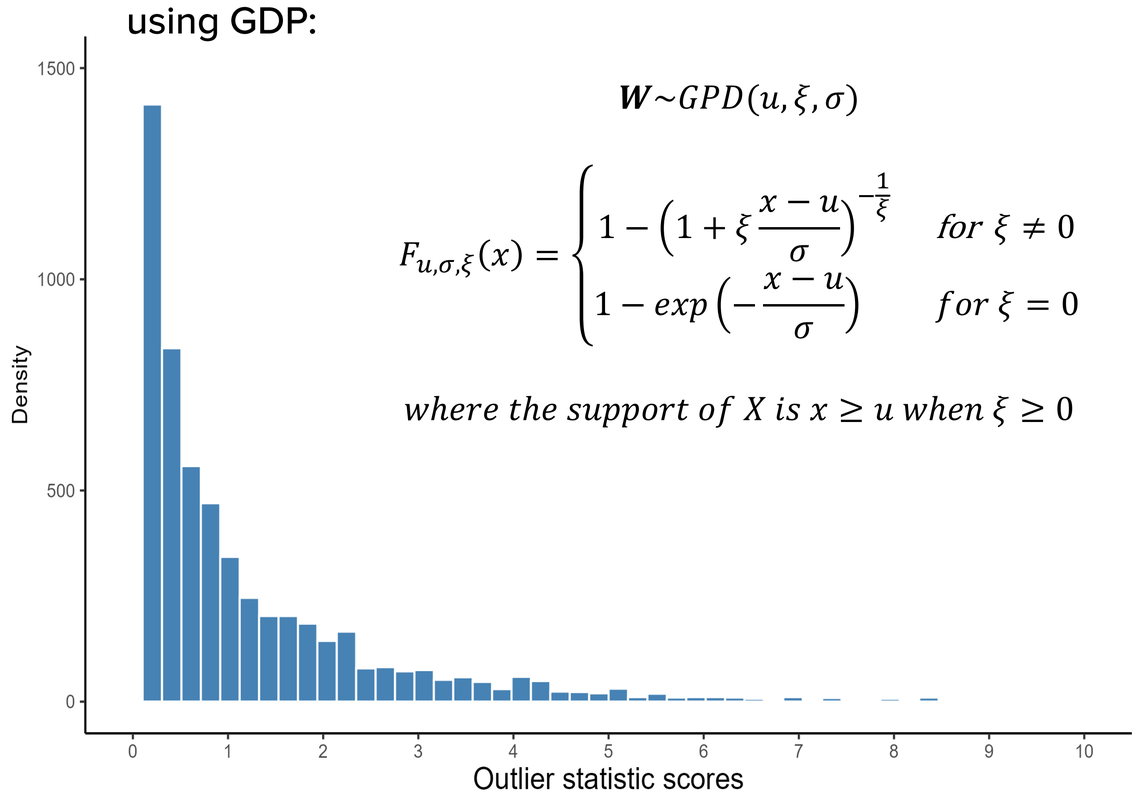
- We are developing **a suit of Extreme Value Theory (EVT)-based methods for improved marker-trait detection.**
- We will model extreme upper tail of trait-marker association outlier scores by using threshold exceedances to fit generalized Pareto distribution (GPD) and estimate tail probabilities for efficient and accurate thresholding in GWAS.
- Correlation among the outlier statistics due to linkage disequilibrium and mixed-model analyses will be accounted for using EVT tools such as declustering, extremal index estimation and multivariate tails.
- The performance of the new approach will be benchmarked against resampling threshold method of Smith & Cullis (2021), Bonferroni and FDR-based thresholds, and accuracy analysis using simulated known genetic architectures.

Given M genetic markers, a vector of marker-trait associations outlier statistic scores $\boldsymbol{x} = (x_1, x_2, \dots, x_M)$ obtained from `wgaim`, and a threshold value u, we define exceedances over threshold $\boldsymbol{W} := \boldsymbol{x} - u | \boldsymbol{x} > u$ and model $\boldsymbol{W}$ using GDP:

$$\boldsymbol{W} \sim GPD(u, \xi, \sigma)$$

$$F_{u,\sigma,\xi}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x-u}{\sigma}\right)^{-\frac{1}{\xi}} & for\ \xi \neq 0 \\ 1 - exp\left(-\frac{x-u}{\sigma}\right) & for\ \xi = 0 \end{cases}$$

$$where\ the\ support\ of\ X\ is\ x \geq u\ when\ \xi \geq 0$$



References

Beirlant, J., Goegebeur, Y., Segers, J. & Teugels, J. (2004) Statistics of Extremes: Theory and Applications. Wiley, Chichester.

de Haan, L. & Ferreira, A. (2006) Extreme Value Theory: An Introduction. Springer, New `York.

Marvin N.W. & Zigler A. (2017) A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77:1-17.

Verbyla, A.P., Cullis, B.R., Thompson, R. & Smith, A.B. (2007) Whole-genome QTL analysis using mixed models and forward selection. Genetics Research, 89:507-514.

Verbyla, A.P., Taylor, J.D. and Verbyla, K.L. (2012) RWGAIM: An efficient high dimensional random whole genome average (QTL) interval mapping approach. Genetics Research, 94: 291–306.

Smith, A. & Cullis, B. (2021) An Efficient Resampling Scheme for Outlier Detection in Linear Mixed Models, Technical Report 18:21, UOW.

Industry impact

Developed methods will provide computational efficiency and statistical precision to meet industry's needs of utilising high-density genotyping and large-scale phenotyping, including remote sensing phenotypes and rapid-cycle breeding. Their performance will be demonstrated against the legacy methods using real breeding and pre-breeding Australian grain datasets.

The support of the Grains Research and Development Corporation (GRDC) provided through the Analytics for the Australian Grains Industry (AAGI) project is acknowledged.