UWA Applied Bioinformatics Group

Mitchell Bestry[1], Shriprabha R Upadhyaya[1], Jack Bruton[1], Tessa R MacNish[1], Teng Li[1], Ze Wu[1], Nimesha Ranawaka[1], Shameela Mohamedikbal[1], Felipe E Albornoz[1], Thomas Bergmann[1], David Edwards[1]

1 Centre for Applied Bioinformatics, The University of Western Australia, Perth, Australia
For enquiries, contact dave.edwards@uwa.edu.au

QR code for online version

Predicting crop traits with machine learning

Machine learning can capture additive, dominance and epistatic effects that influence a trait, often modelling these relationships more effectively than traditional methods. In genomic selection, machine learning models are used to associate genetic markers with crop traits through feature importance analyses. This can improve the accuracy and efficiency of selecting superior genotypes.

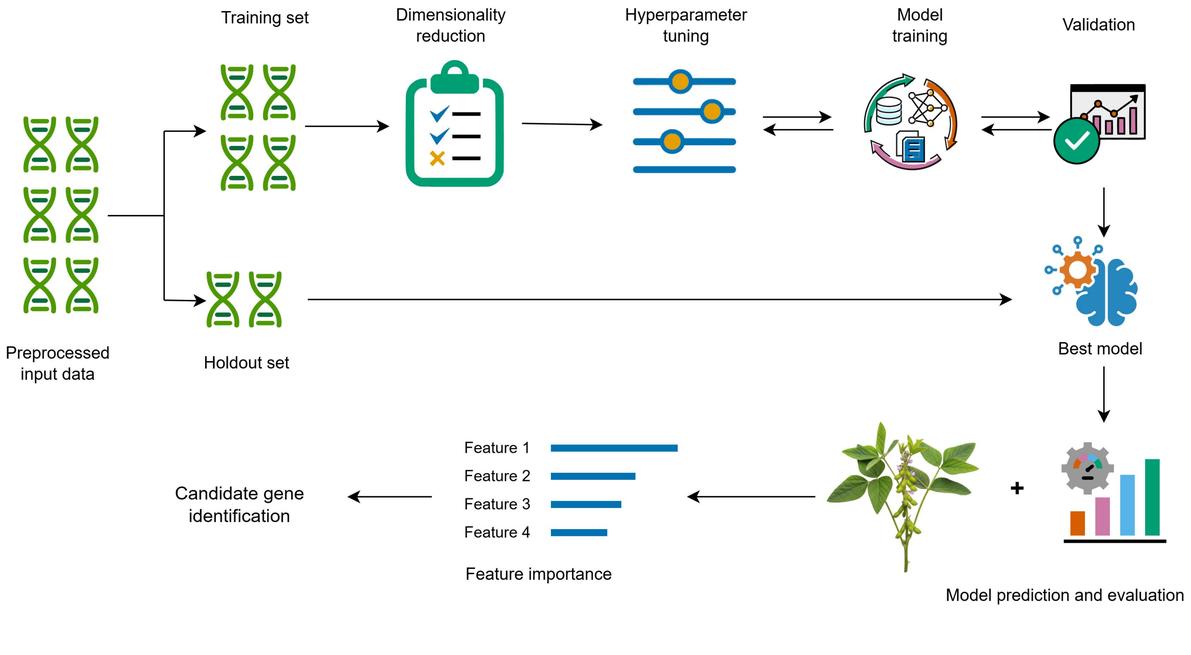


Fig 1: Flowchart of a machine learning pipeline for trait prediction.

Extracting crop data with large language models

Extracting large volumes of crop data from documents is a time-consuming process. Large language models (LLMs) provide a mechanism by which these tasks can be executed automatically with high accuracy. We are building LLM-based data extraction pipelines, processing thousands of publications, documents and tables to generate structured datasets. These datasets are being used to build decision support applications for crop breeders, growers and researchers.

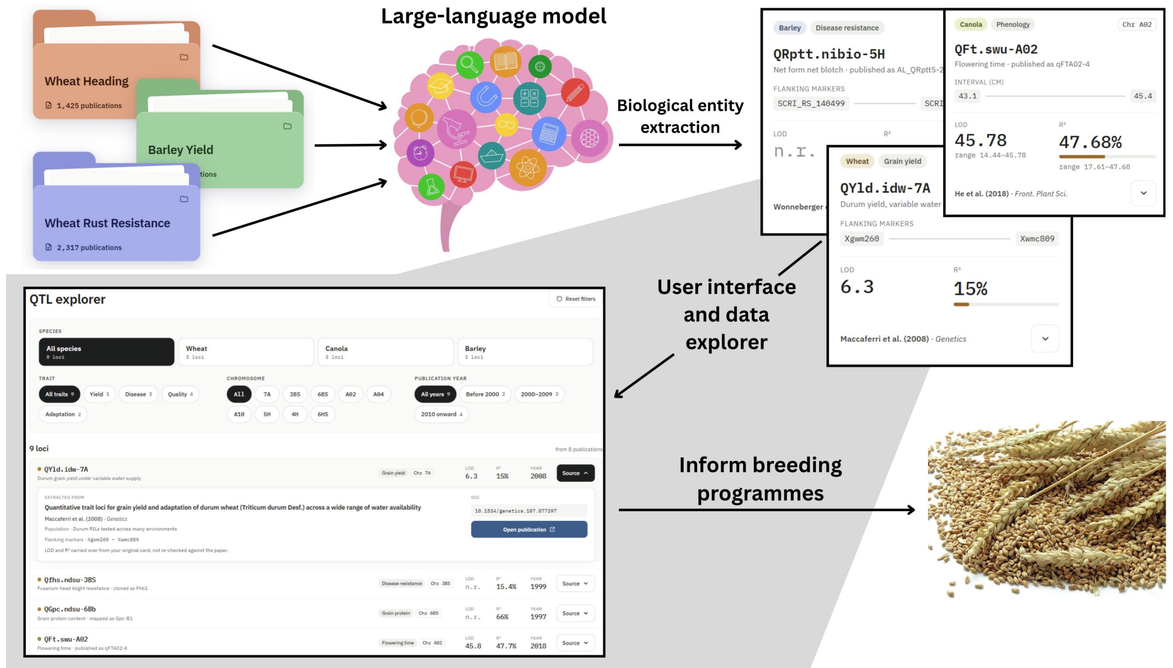


Fig 2: Diagram outlining the LLM-based data extraction pipeline.

Defining haplotypes in crops to optimise breeding

Haplotypes can improve the accuracy and efficiency of genomic selection while also capturing potentially novel genetic variation and broadening the diversity of alleles for selection. Characterising alleles allows breeders to identify which haplotype variants have deleterious, negligible, or beneficial effects on their crops. We have developed HaploVar (1), an R package designed to support haplotype-based genomic selection.

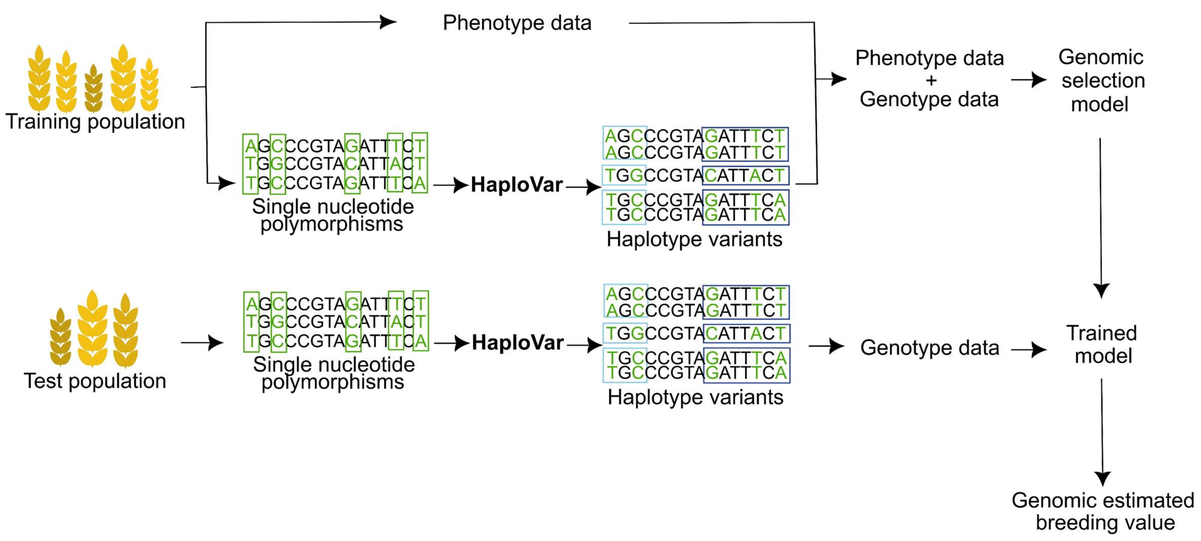


Fig 3: Flowchart outlining the application of HaploVar in genomic selection.

Optimising yields with crop image analysis

We have applied artificial intelligence to aerial crop images for optimising yields. In one example, we trained deep learning models to identify canola against other plants including weeds in aerial plot images (2). This could enable targeted weed control with higher effectiveness and reduced herbicide usage.

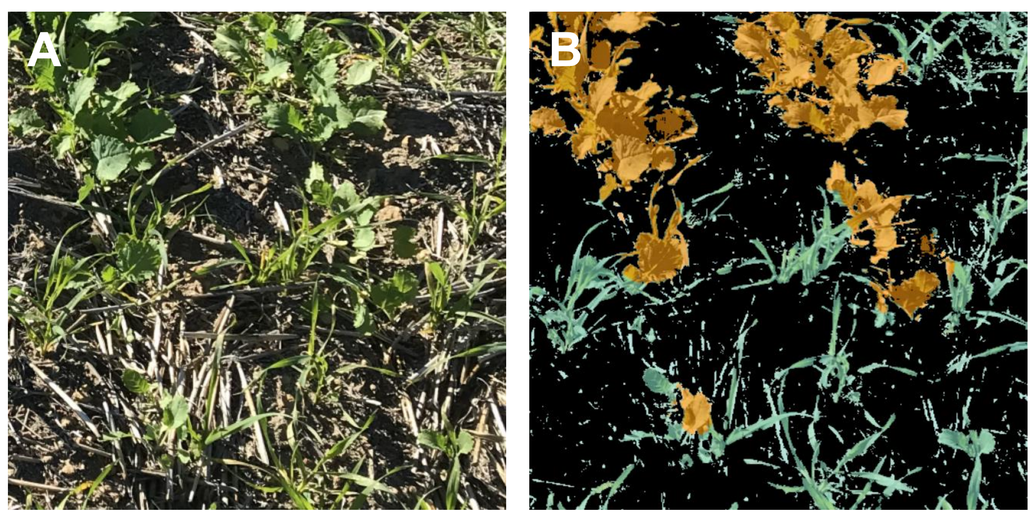


Fig 4: Canola identification in a field image: (A) RGB image and (B) deep learning model prediction, with canola shown in orange and all other plants, including weeds, in teal (2).

Capturing genetic variation in crops with pangenomes

Single reference genomes do not capture the extensive presence/absence variation observed between individuals of the same plant species. Pangenomes can capture this diversity. We have assembled iterative and graph-based pangenomes for multiple crops including wheat (3) and canola (4). As pangenomes provide a broader representation of species genome content, they capture more heritability than single references, improving the accuracy of trait association and genomic selection.

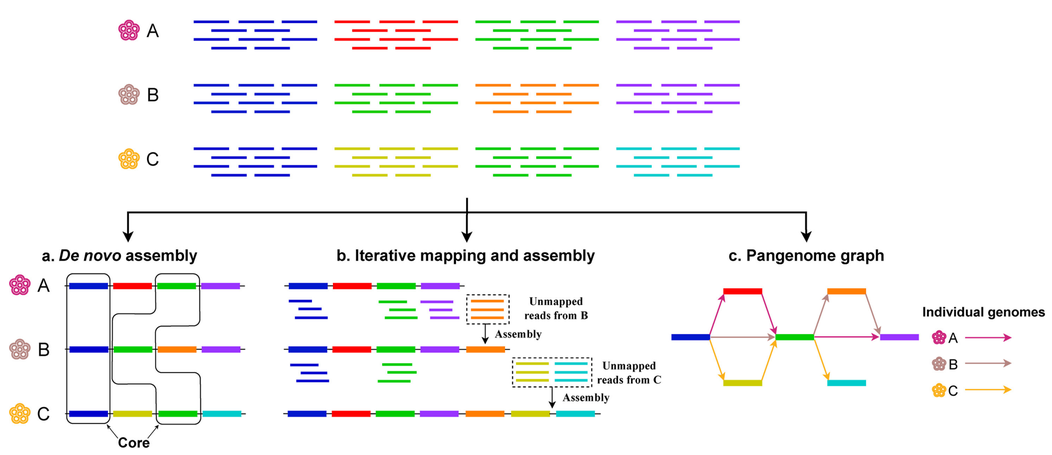


Fig 5: Schematic diagram illustrating three approaches (a, b and c) for pangenome construction from three genomes (A, B and C). Reads from shared genomic segments are indicated with the same colours. Adapted from (5).

Identifying gene presence/absence variation in wheat

Pangenome analysis has shown that as many as 40% of wheat genes are absent from some varieties. We analysed 827 geographically diverse landraces and identified 10,426 high-confidence annotated genes that are absent from all 217 modern cultivars studied (6). This untapped gene reservoir is enriched with genes for biotic and abiotic stress and represents a valuable resource for wheat improvement.

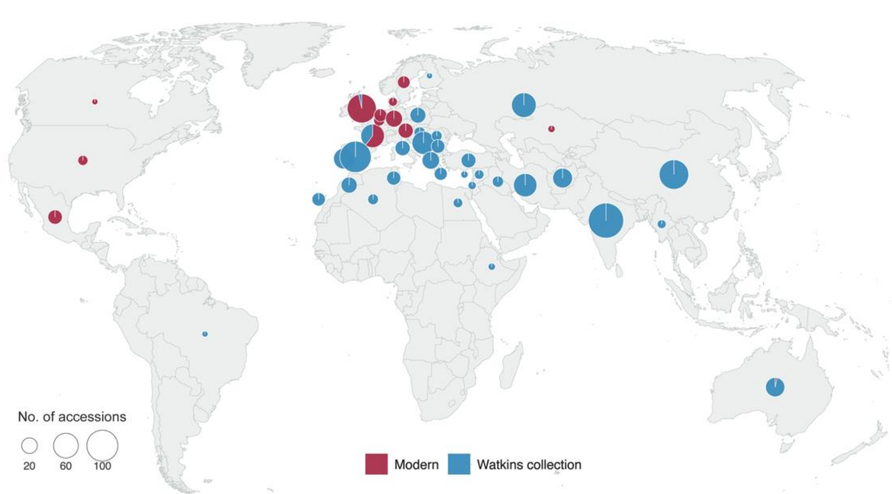


Fig 6: Map showing the geographical distribution of Watkins landraces (blue) and modern cultivars (red) used in this study (6).

References

1. MacNish TR, Al-Mamun HA, Bergmann T, Bestry MS, Marsh JI, Edwards D. HaploVar: An R package for defining local haplotype variants for trait association and trait prediction analyses. Bioinformatics. 2025;41(12):btaf602.
2. Mckay M, Danilevicz MF, Ashworth MB, Rocha RL, Upadhyaya SR, Bennamoun M, et al. Focus on the Crop Not the Weed: Canola Identification for Precision Weed Management Using Deep Learning. Remote Sens. 2024;16(11):2041.
3. Bayer PE, Petereit J, Durant É, Monat C, Rouard M, Hu H, et al. Wheat Panache: A pangenome graph database representing presence–absence variation across sixteen bread wheat genomes. Plant Genome. 2022;15(3):e20221.
4. MacNish TR, Al-Mamun HA, Bayer PE, McPhan C, Fernandez CGT, Upadhyaya SR, et al. Brassica Panache: A multi-species graph pangenome representing presence absence variation across forty-one Brassica genomes. Plant Genome. 2025;18(1):e20535.
5. Li T, Mohamedikbal S, Bestry M, Batley J, Edwards D. Pangenomics combined with artificial intelligence and precision breeding can accelerate crop improvement. Curr Opin Plant Biol. 2025;88:102825.
6. Li T, Albornoz FE, Bruton J, Bestry M, Dolatabadian A, MacNish T, et al. Capturing the missing gene content of elite wheat by analysis of wheat landraces. bioRxiv. 2026:2026.07.30.741942.