#129

Using pangenomics and machine learning for disease resistance

David Edwards[1]

[1] Centre for Applied Bioinformatics, University of Western Australia, Australia

As more genome sequences were assembled for Brassicas, it became obvious that there is significant gene presence/absence variation between individuals, with as many as 40% of *B. napus* genes being absent in one or more individuals[1,2]. The predicted functions of these variable genes are often associated with biotic and abiotic stress, and so knowledge of their presence or absence is useful for breeding varieties with enhanced adaptability and disease resistance. Pangenomes have been constructed for most crop species to capture the variability in gene presence between individuals. We have constructed several Brassica pangenomes[3], the latest graph pangenome incorporating individuals from *B. oleracea* and *B. rapa*, and these highlight the variation in disease resistance genes between individuals.

The explosion of available genetic and trait data has opened opportunities to apply machine learning (ML) approaches to mine this data to find associations between heritable traits and genetic variation[4]. These approaches are flexible - accepting image, text, and tabular data - with the ability to combine outputs for powerful merged models. We have developed multi modal ML models that can incorporate diverse trait data to predict quantitative and qualitative crop traits[5,6]. These include the presence of specific R genes in canola based on phenotypic assays, the level of quantitative resistance to blackleg infection based on canola genotype data and the prediction of plant stress using drone image data[7].

There remains significant scope to extend these approaches with the continued growth of available data. For example, the recent success of large language models (LLMs) such as ChatGPT has highlighted the capability and limitations of text mining. We have developed a preliminary LLM model based on ~1000 recent publications relating to crop disease, and while there remain challenges to develop such tools for broad use, the ability to integrate text data into prediction and discovery models offers significant long-term potential.

References:

1. Hurgobin, B. *et al*. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. (2018). *Plant Biotechnology Journal* **16**, 1265-1274
2. Golicz, A.A., Batley, J. & Edwards, D. (2016). Towards plant pangenomics. *Plant Biotechnology Journal* **14**, 1099-1105.
3. Bayer, P.E. *et al*. (2021). Modelling of gene loss propensity in the pangenomes of three Brassica species suggests different mechanisms between polyploids and diploids. *Plant Biotechnology Journal* **19**, 2488-2500
4. Bayer, P.E. et al. (2021). The application of pangenomics and machine learning in genomic selection in plants. Plant Genome **14**, e20112
5. Danilevicz, M.F., Bayer, P.E., Boussaid, F., Bennamoun, M. & Edwards, D. (2021). Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection. *Remote Sensing* **13**, 3976.
6. Gill, M. *et al*. (2022). Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. *BMC Plant Biology* **22**, 180.

Khotimah, W.N. *et al*. (2023). MCE-ST: Classifying crop stress using hyperspectral data with a multiscale conformer encoder and spectral-based tokens. *International Journal of Applied Earth Observation and Geoinformation* **118**, 103286.